Applied Multivariate Analysis

Myung Hee Lee

Fall 2014

Tentative Schedule

- introduction
- descriptive techniques
- matrix algebra
- multivariate distribution
- multivariate normal distribution
- principal component analysis
- discriminant analysis
- factor analysis
- cluster analysis
- canonical correlation analysis
- graphical models

Multivariate data set?

- A data set in which several variables are measured on each sampled unit.
- We have n units (individuals) and p variables in a multivariate data set, where $p > 1$.
- Notation:

Visualization of multivariate data?

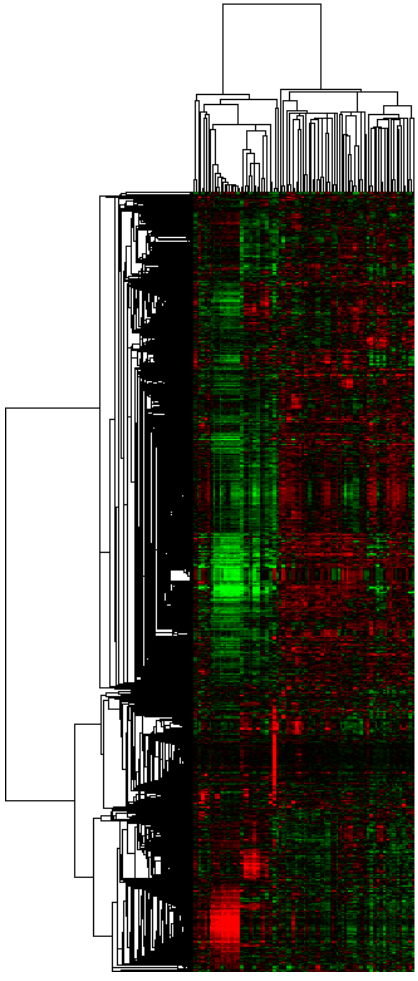
- loosely, one can think of each variables as an additional dimension
- many variable data = data sitting in a high dimensional space
- how do we look at the data?
  - 1-d? histogram
  - 2-d? scatter plot
  - 3-d? 3-D scatter plot? (spinning) point cloud view
  - 100-d?

Visualization of multivariate data?

- loosely, one can think of each variables as an additional dimension
- many variable data = data sitting in a high dimensional space
- how do we look at the data?
  - 1-d? histogram
  - 2-d? scatter plot
  - 3-d? 3-D scatter plot? (spinning) point cloud view
  - 100-d?

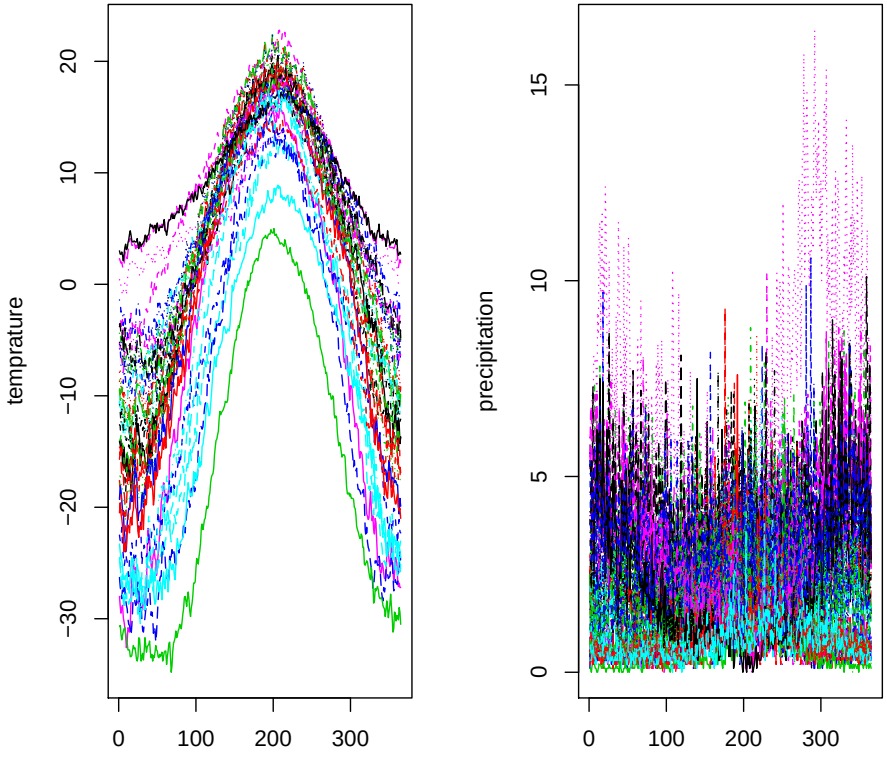
    projection: there are many directions of interest on which projection is useful

Example: Ovarian Cancer Data



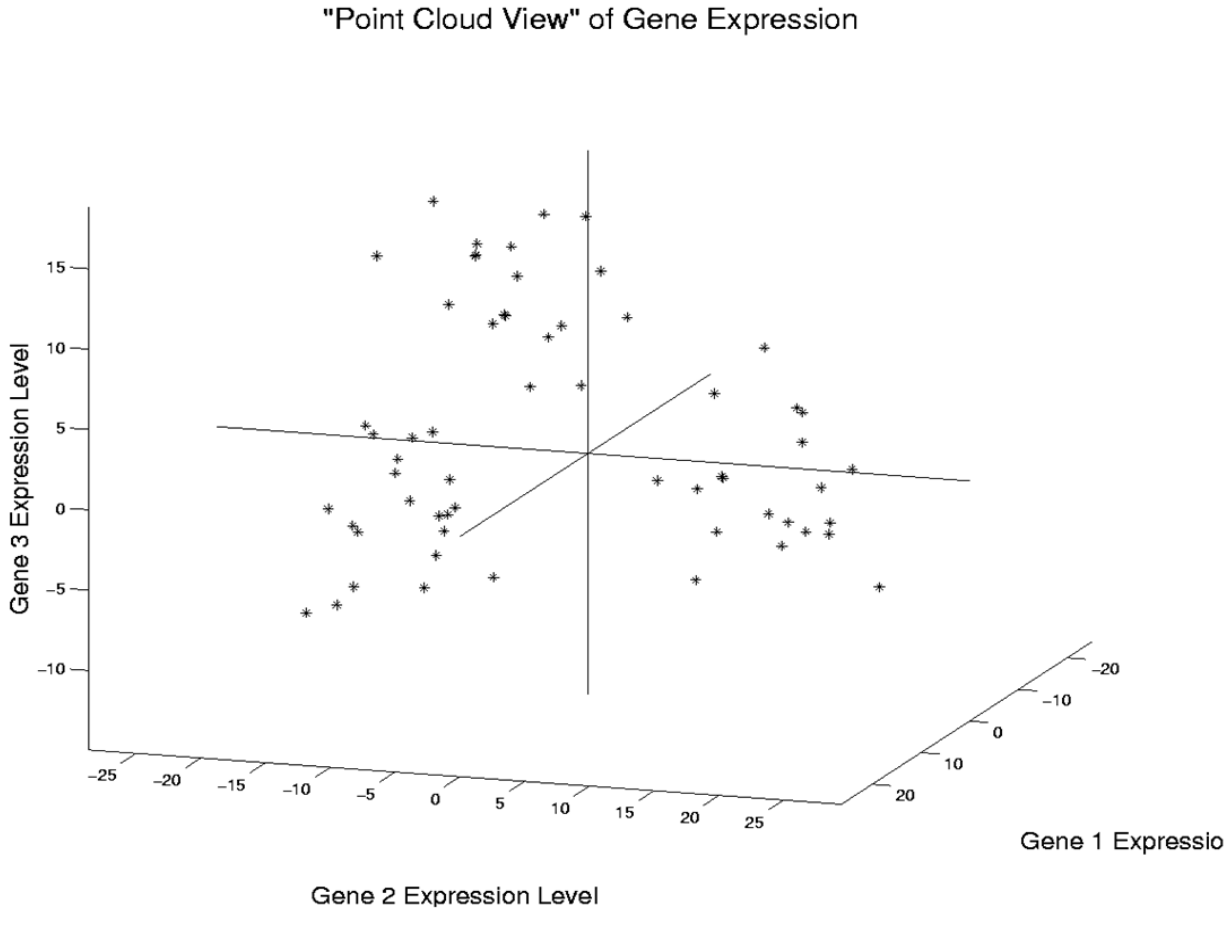
$n = 108$ and $p = 28,475$

Example: Weather data



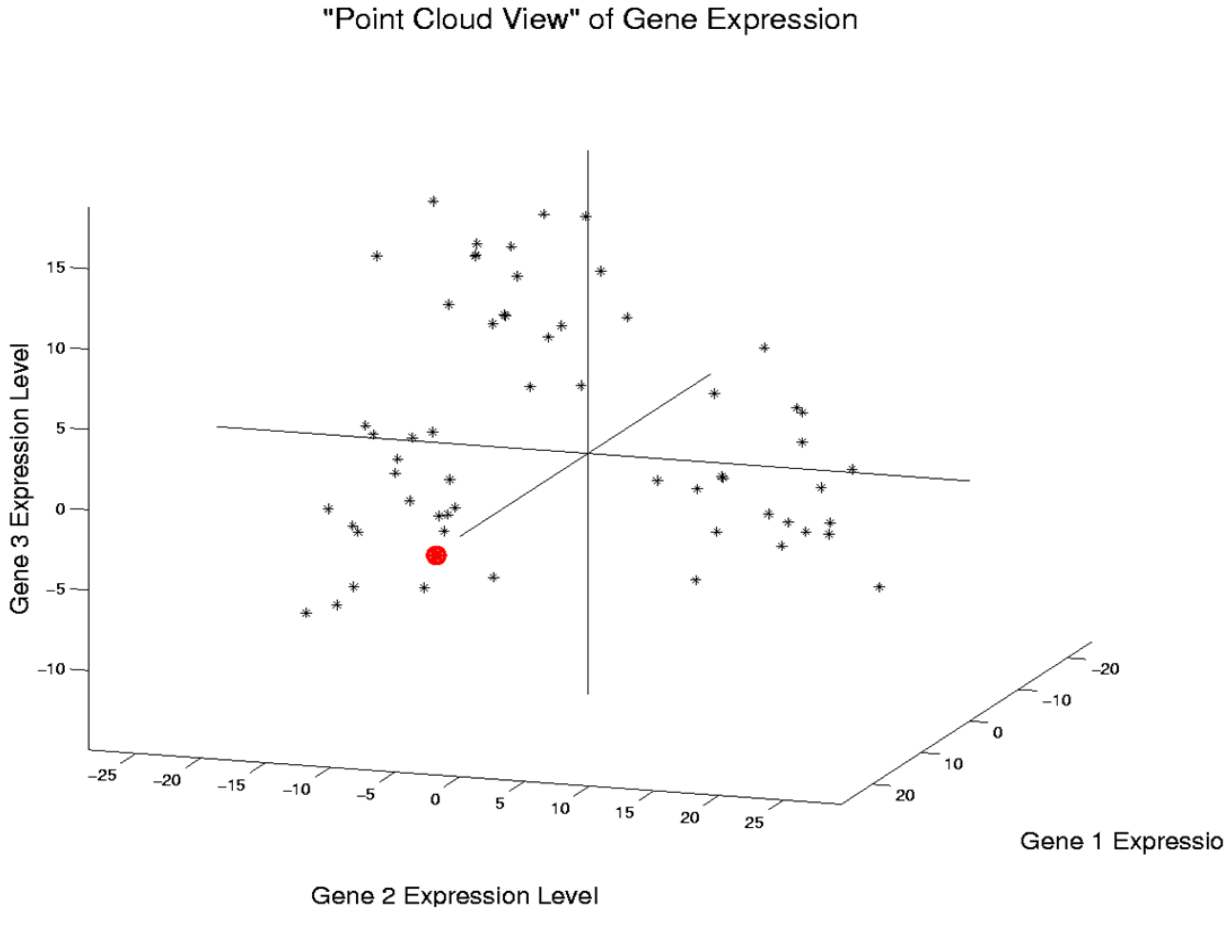
$n = 35$ and $p = 365$

Example

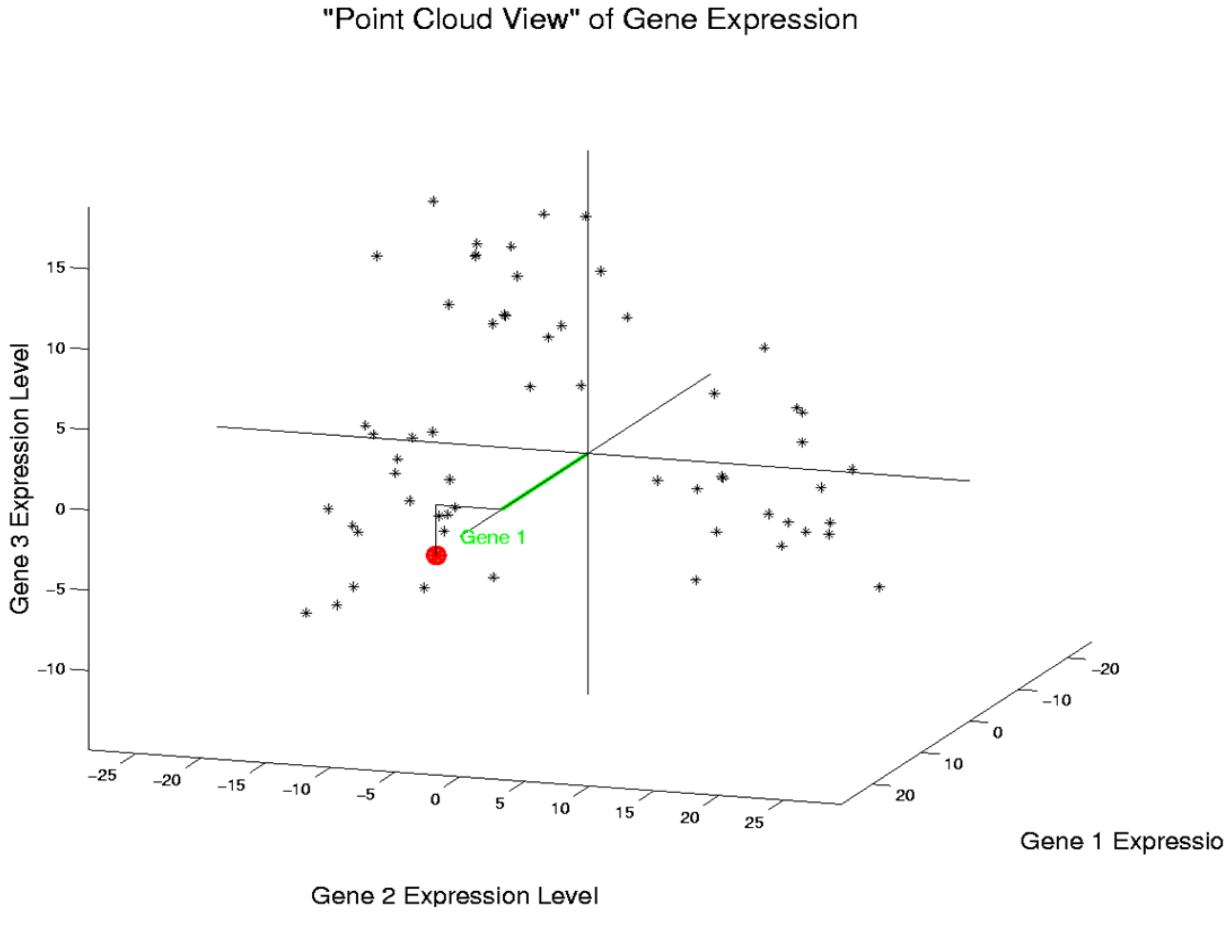


$n = 60$ and $p = 3$

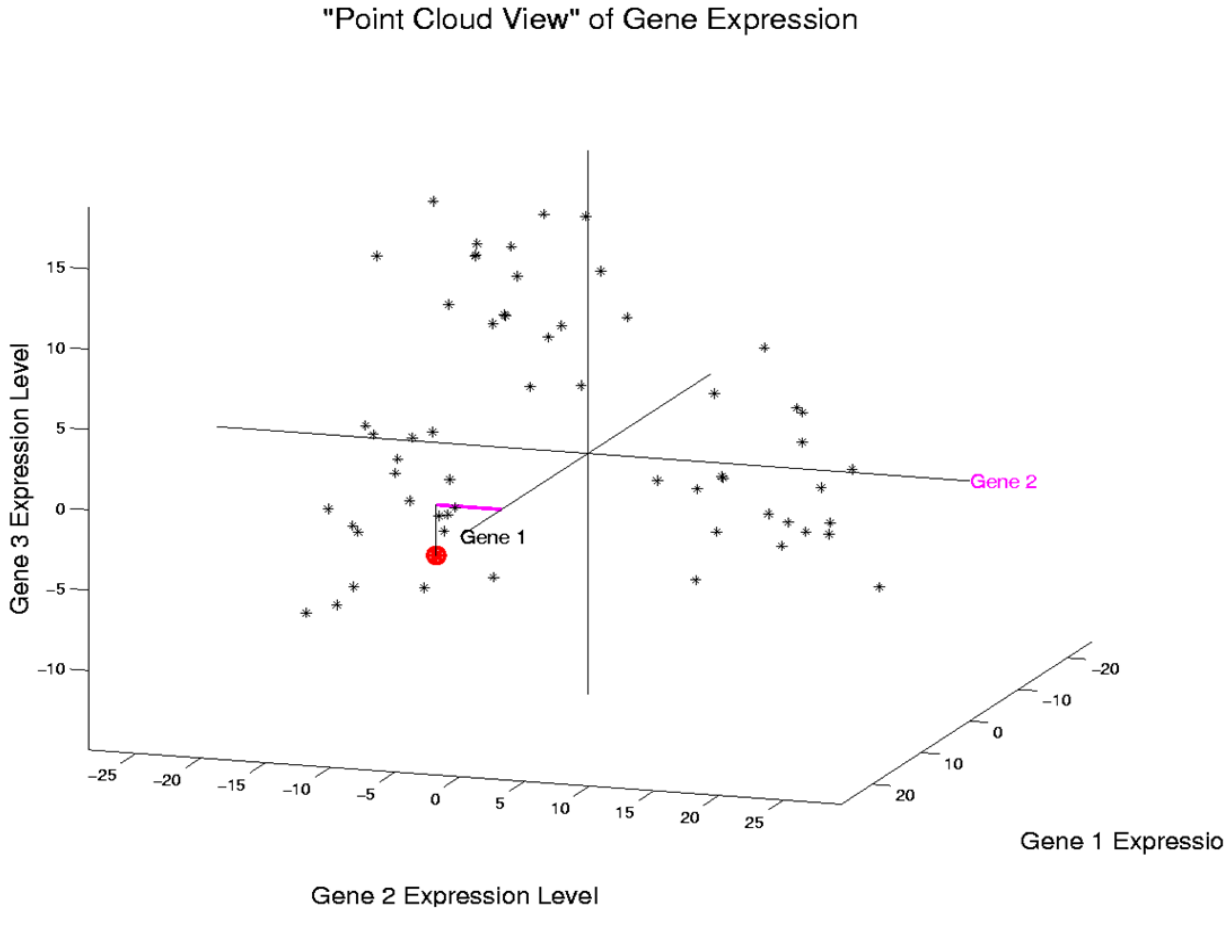
Example


"Point Cloud View" of Gene Expression
Gene 3 Expression Level
15
10
5
0
−5
−10
−25
−20
−15
−10
−5
0
5
10
15
20
25
Gene 2 Expression Level
−20
−10
0
10
20
Gene 1 Expressio

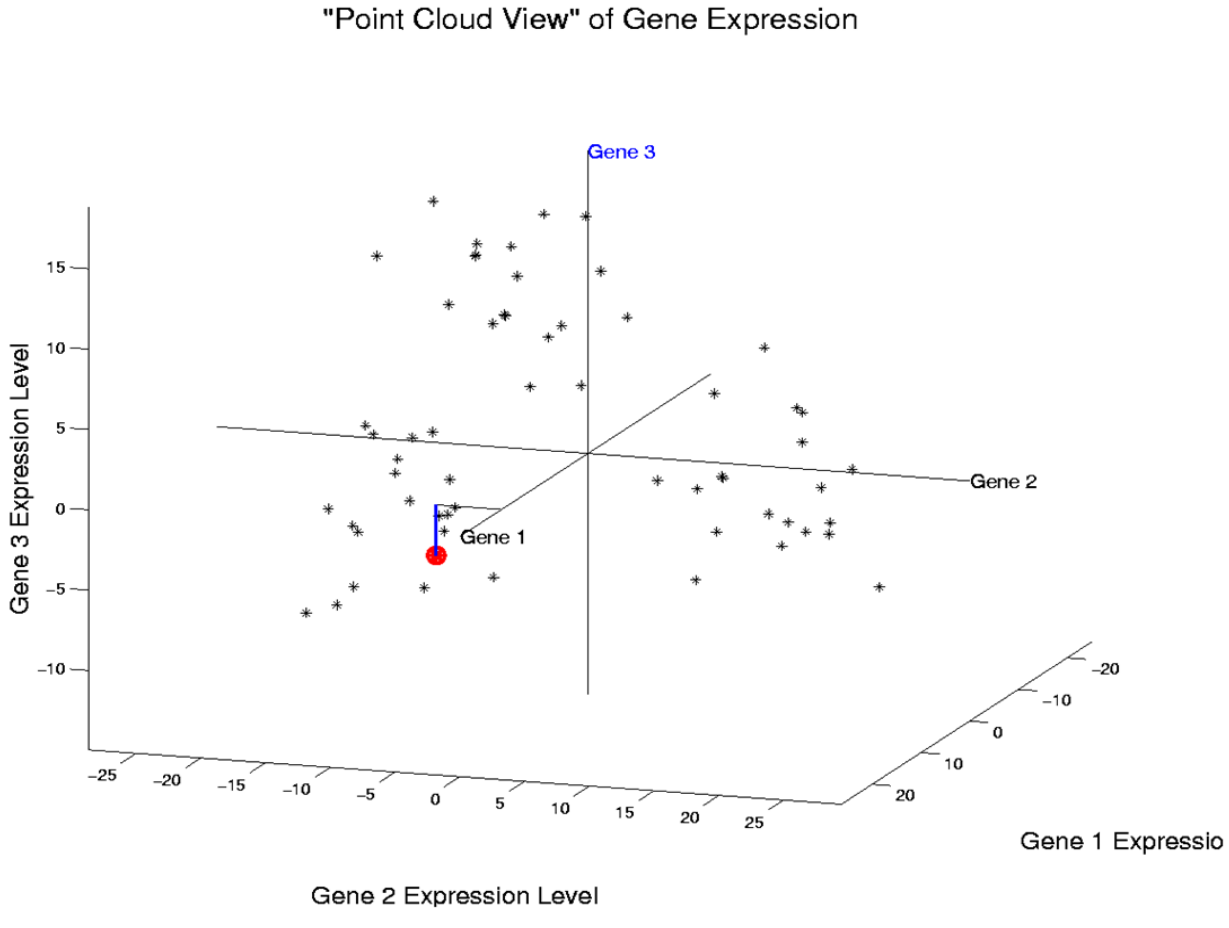
Example


"Point Cloud View" of Gene Expression
Gene 3 Expression Level
15
10
5
0
−5
−10
Gene 1
−25
−20
−15
−10
−5
0
5
10
15
20
25
Gene 2 Expression Level
−20
−10
0
10
20
Gene 1 Expressio

Example

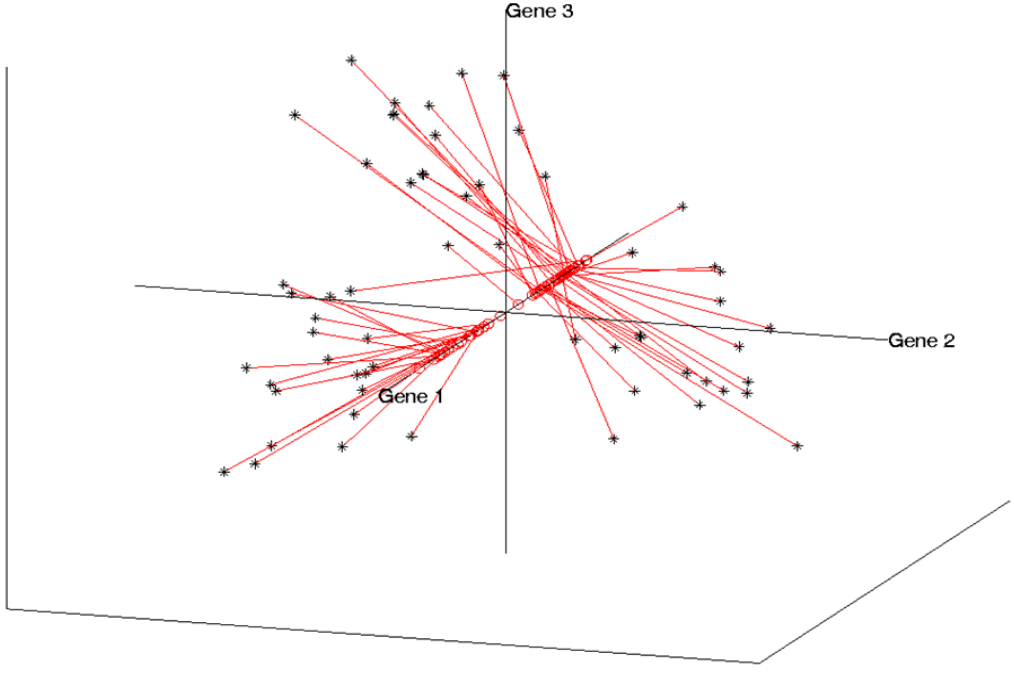

"Point Cloud View" of Gene Expression
Gene 3 Expression Level
15
10
5
0
−5
−10
Gene 2
Gene 1
−25
−20
−15
−10
−5
0
5
10
15
20
25
Gene 2 Expression Level
−20
−10
0
10
20
Gene 1 Expressio

Example

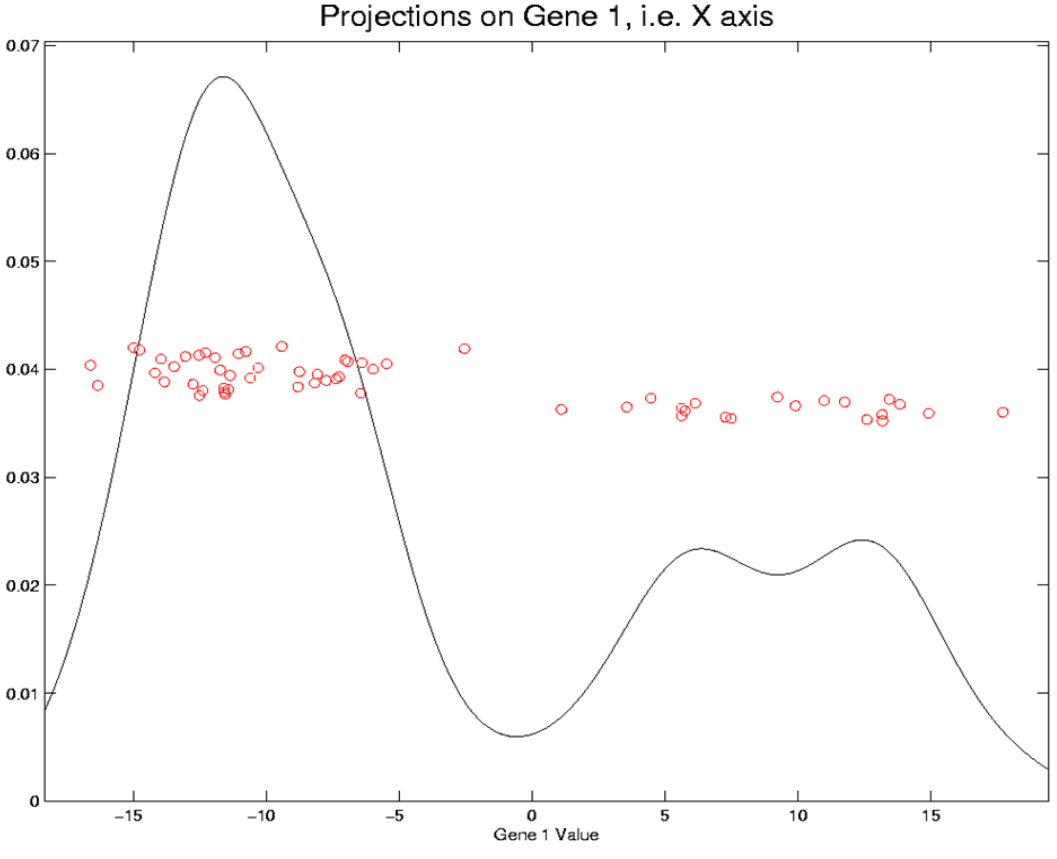

"Point Cloud View" of Gene Expression
Gene 3
Gene 2
Gene 1
Gene 3 Expression Level
Gene 2 Expression Level
Gene 1 Expressio
15
10
5
0
−5
−10
−25
−20
−15
−10
−5
0
5
10
15
20
25
−20
−10
0
10
20

Example

Projections on Gene 1, i.e. X axis

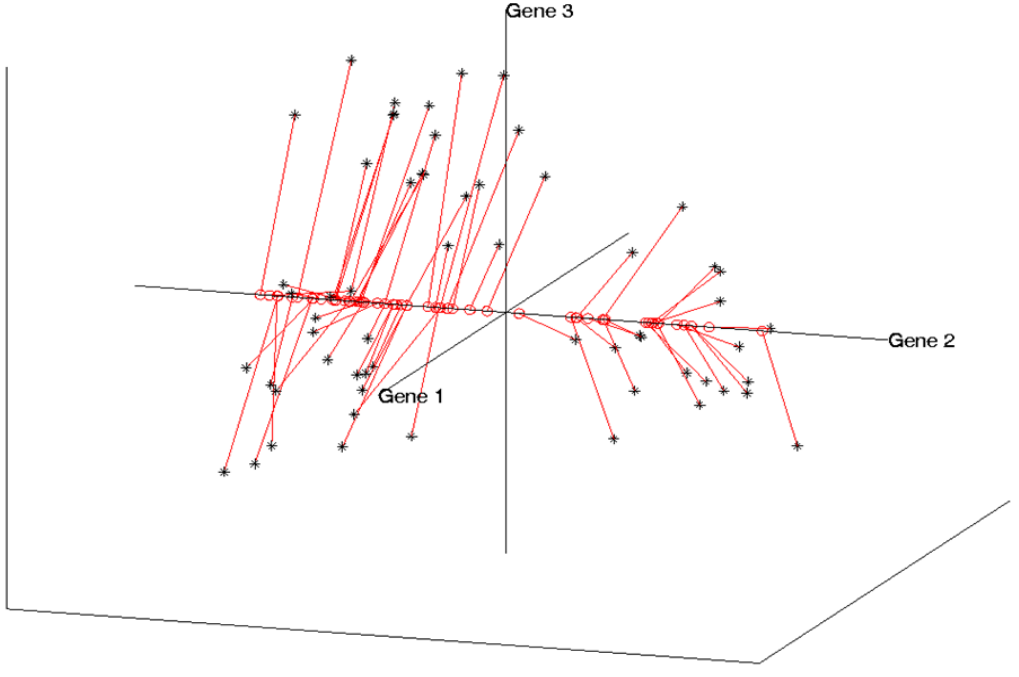


Example

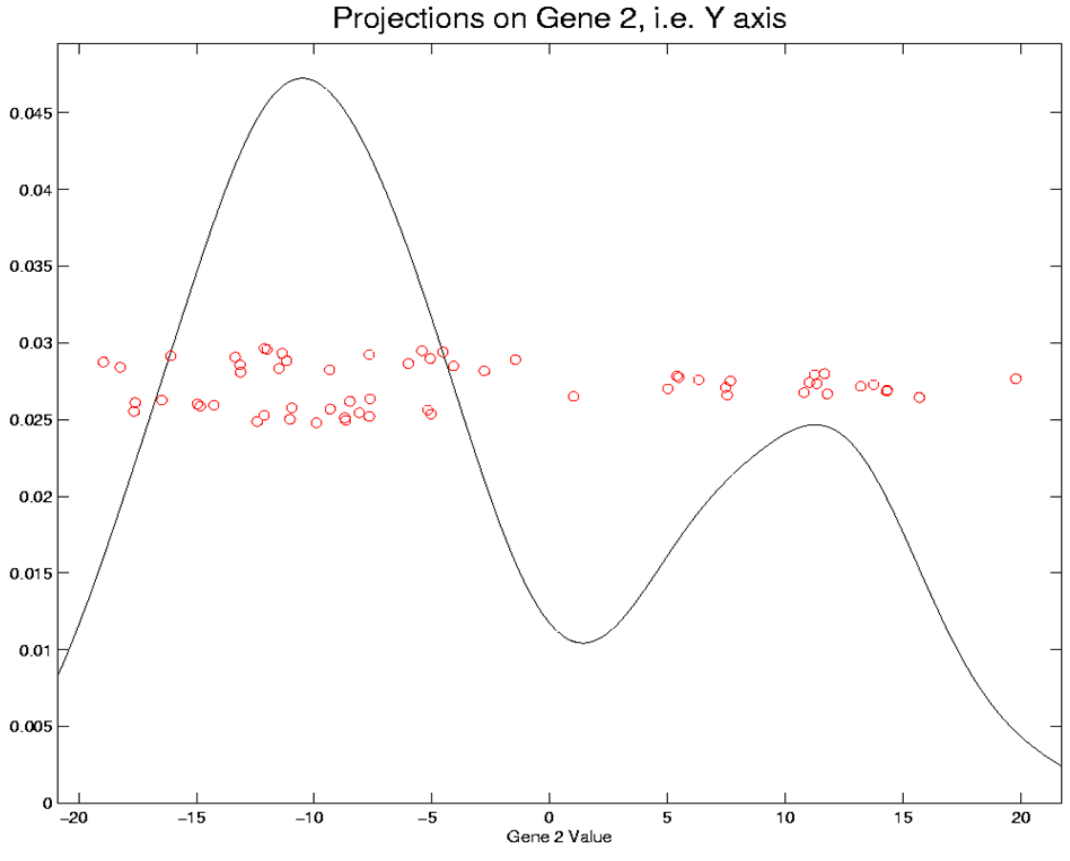

Projections on Gene 1, i.e. X axis
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
−15
−10
−5
0
5
10
15
Gene 1 Value

Example

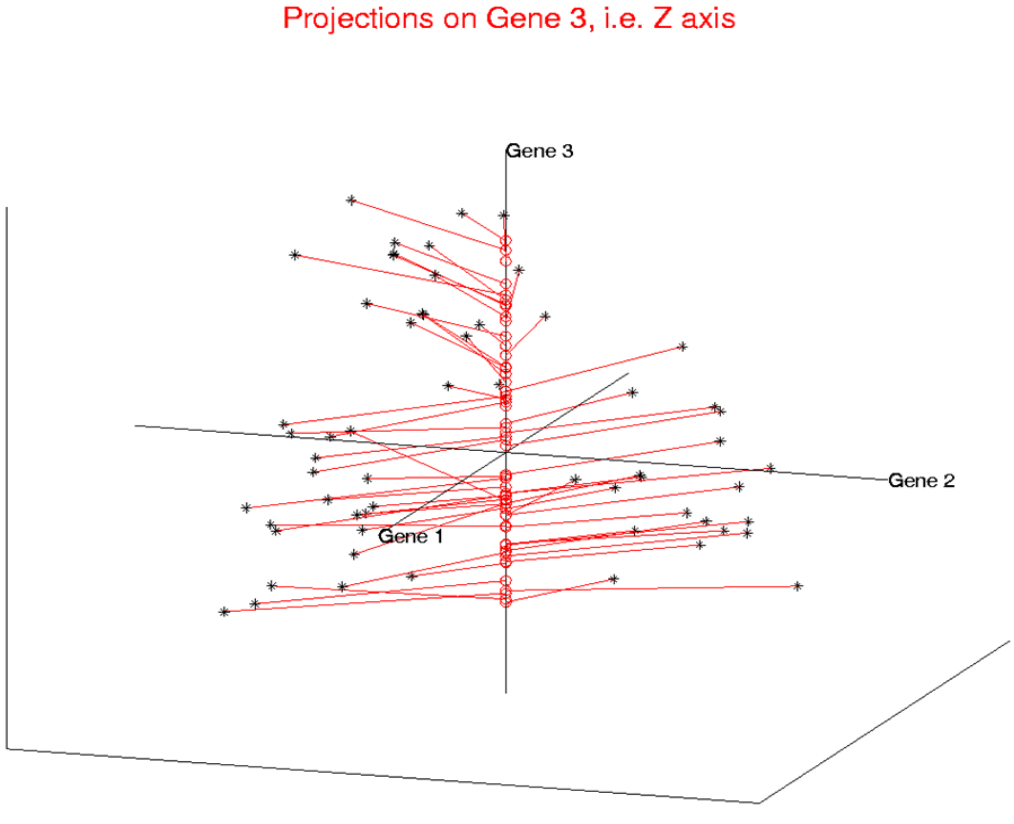
Projections on Gene 2, i.e. Y axis



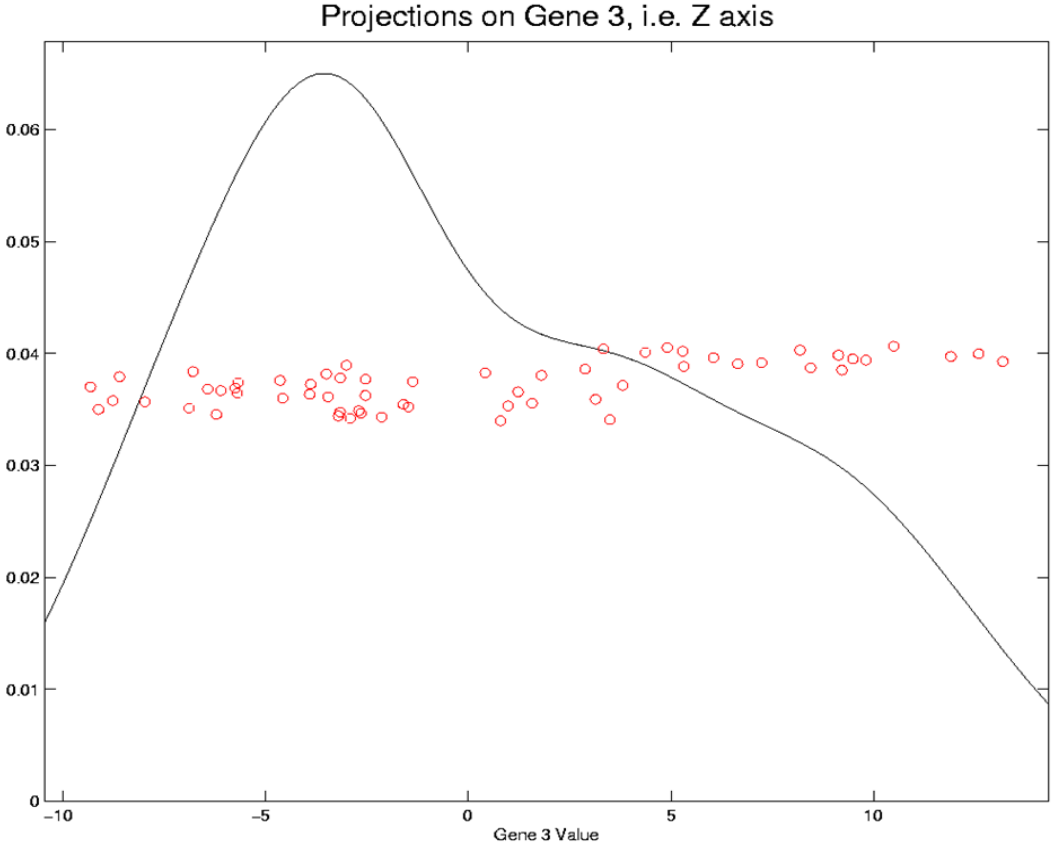
Example


Projections on Gene 2, i.e. Y axis
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
−20
−15
−10
−5
0
5
10
15
20
Gene 2 Value

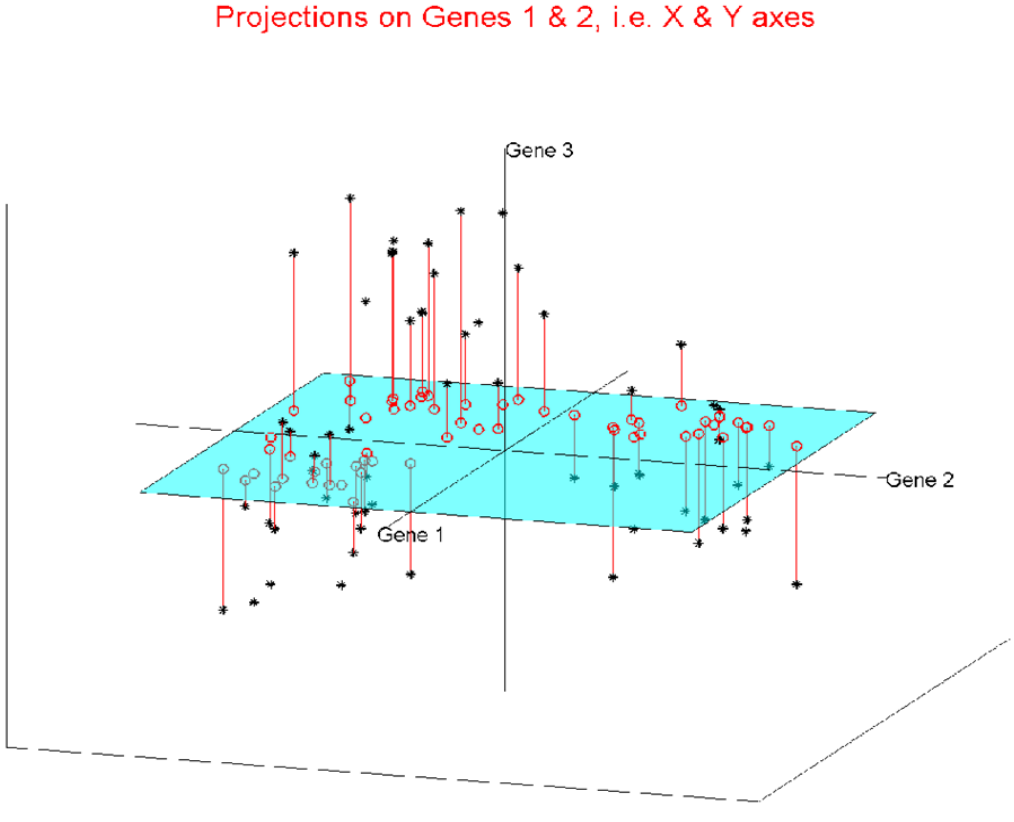
Example



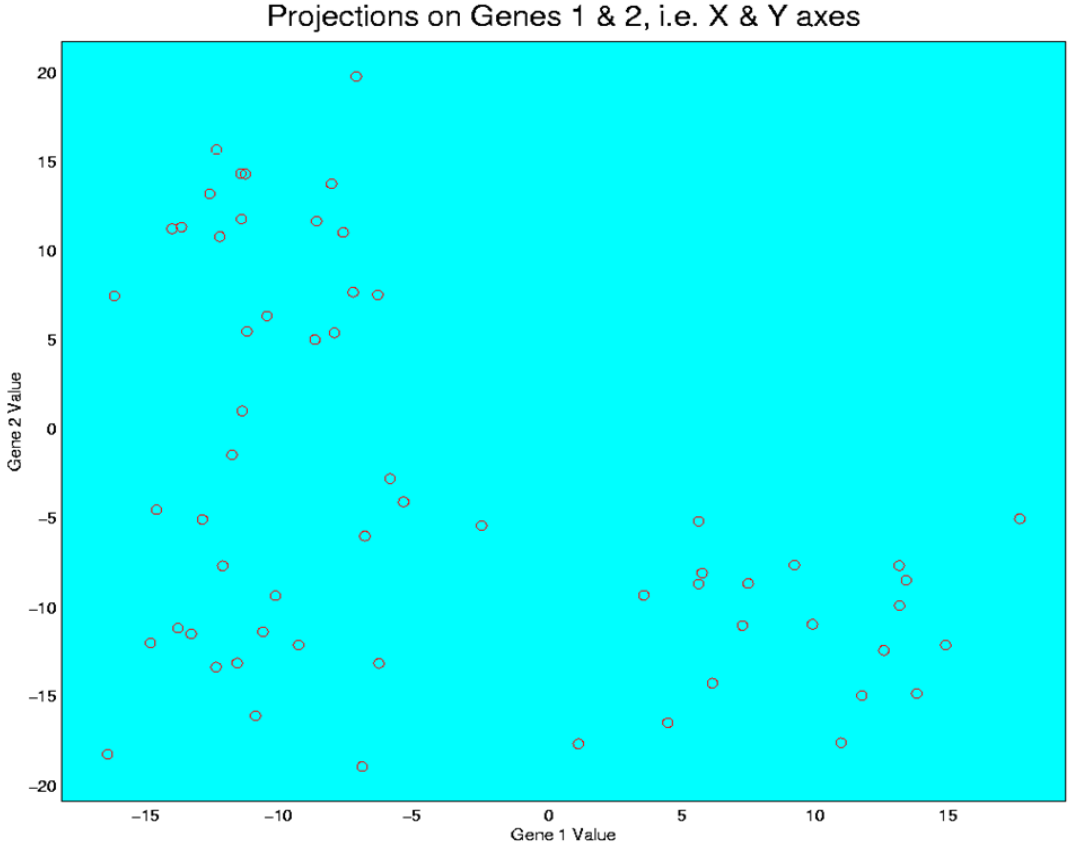
Example


Projections on Gene 3, i.e. Z axis
0.06
0.05
0.04
0.03
0.02
0.01
0
−10
−5
0
5
10
Gene 3 Value

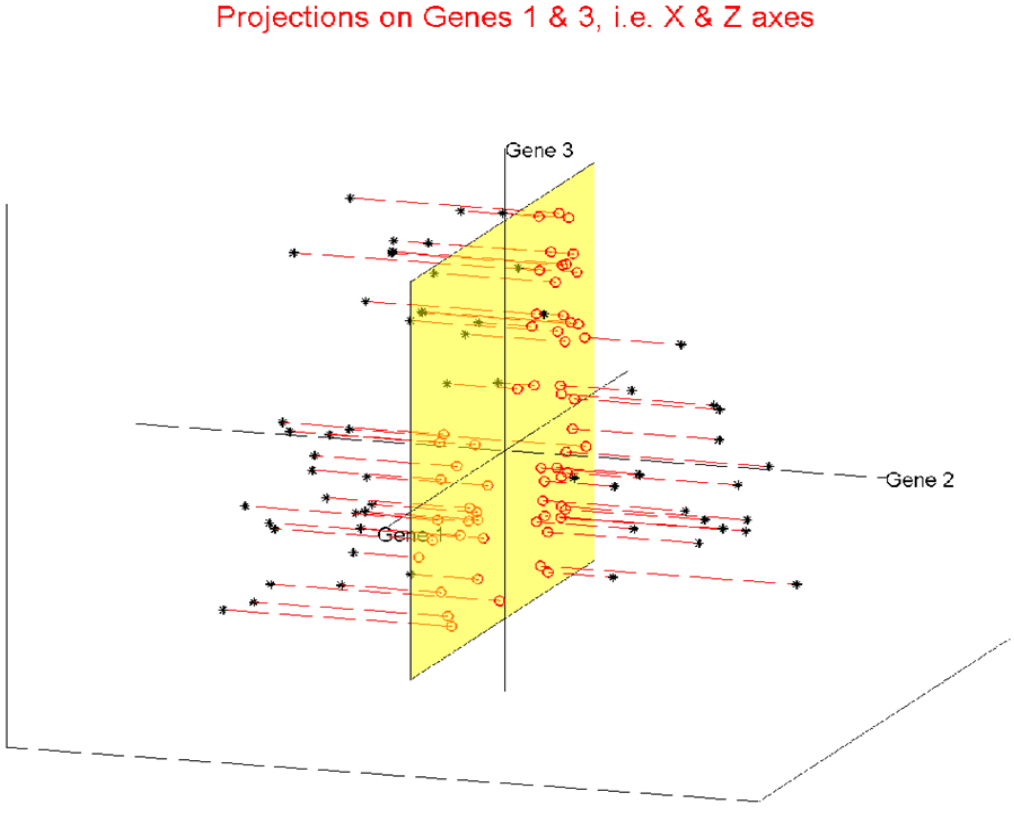
Example



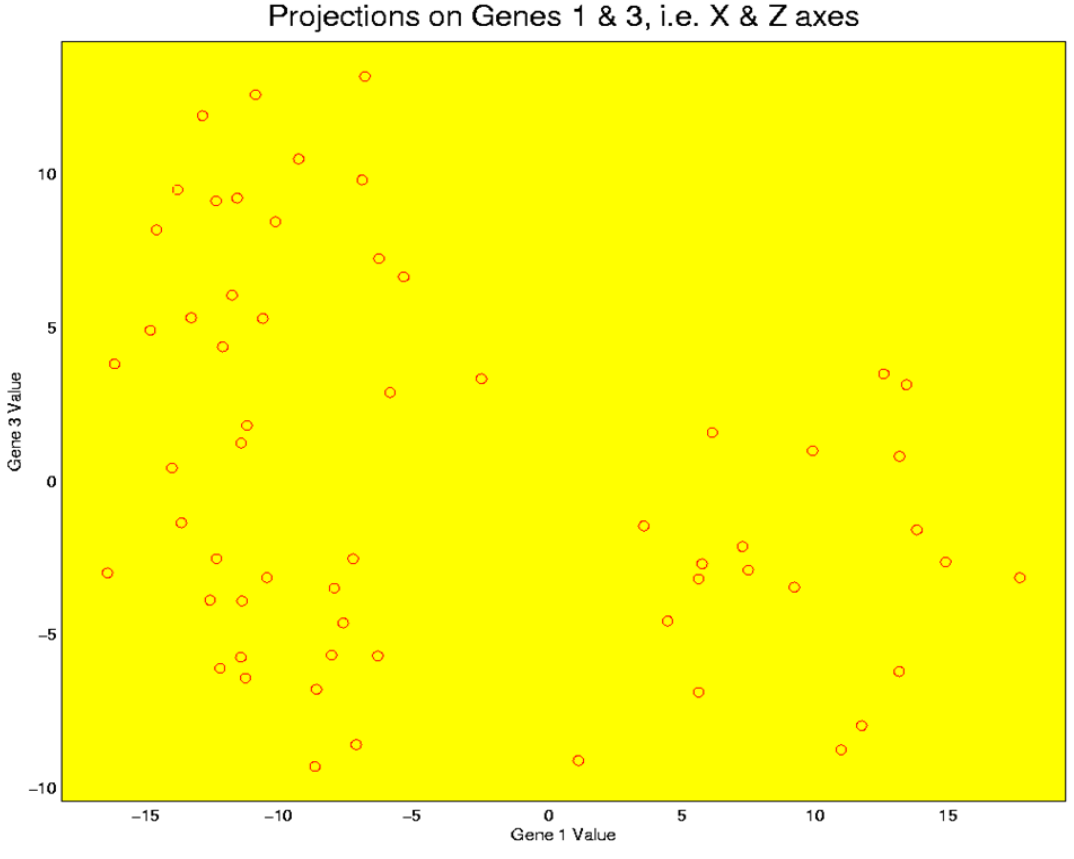
Example


Projections on Genes 1 & 2, i.e. X & Y axes
Gene 2 Value
Gene 1 Value
20
15
10
5
0
−5
−10
−15
−20
−15
−10
−5
0
5
10
15

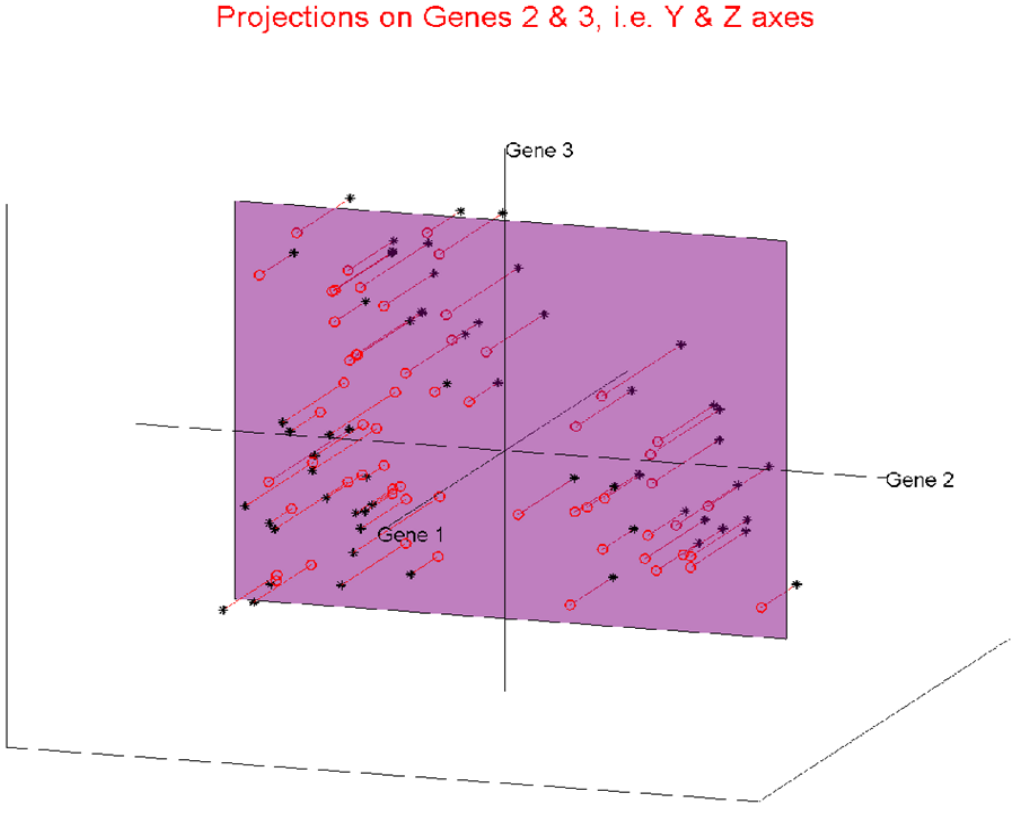
Example



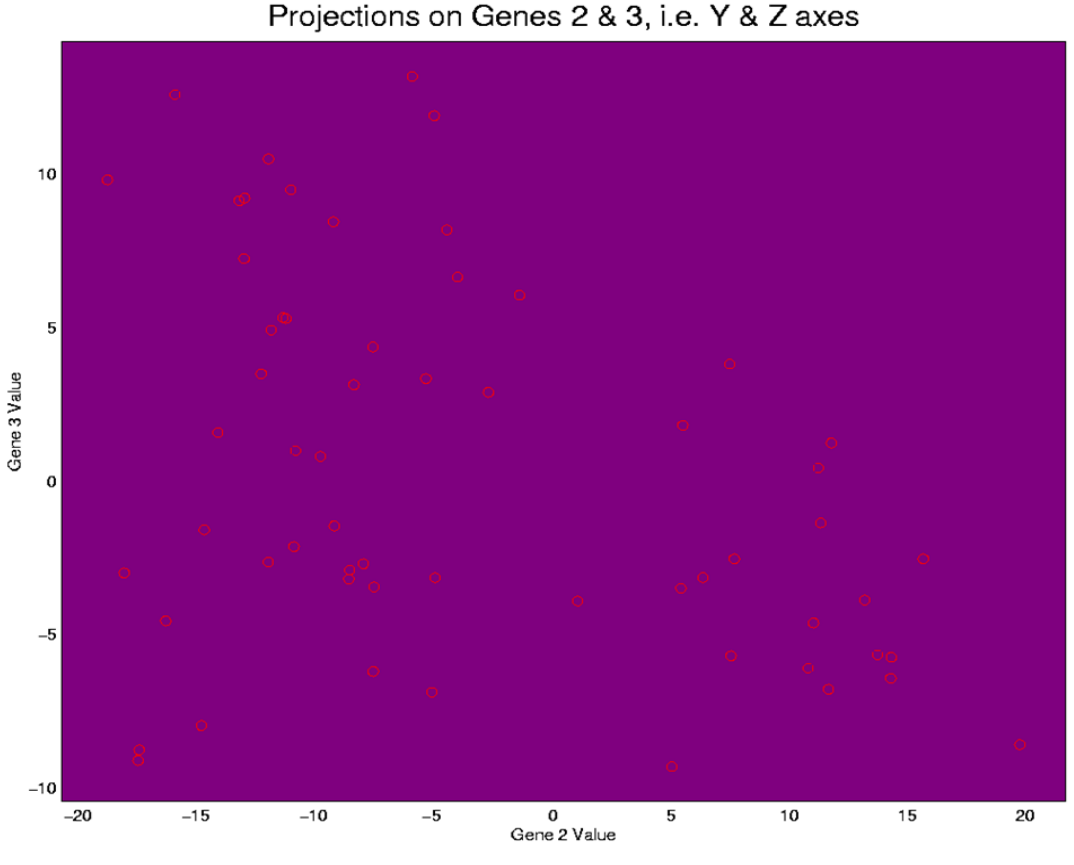
Example


Projections on Genes 1 & 3, i.e. X & Z axes
Gene 3 Value
10
5
0
−5
−10
−15
−10
−5
0
5
10
15
Gene 1 Value

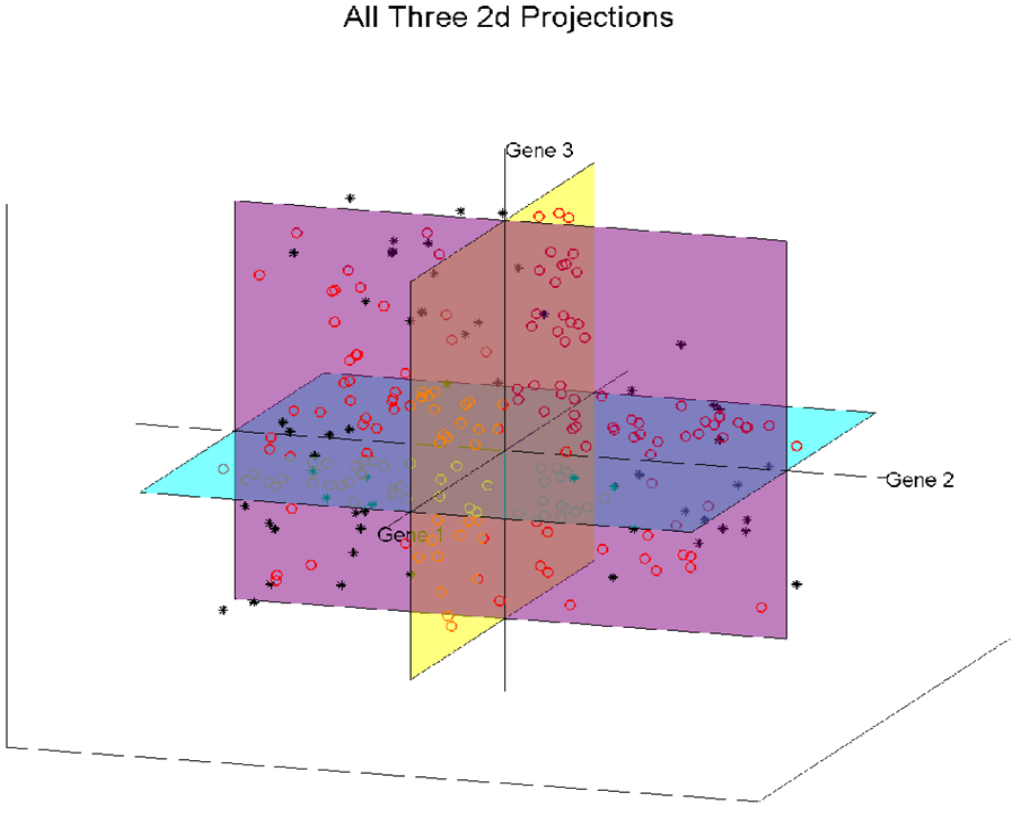
Example



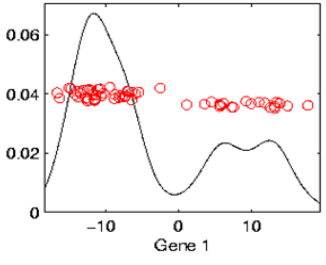
Example

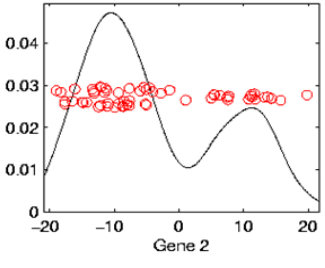

Projections on Genes 2 & 3, i.e. Y & Z axes
Gene 3 Value
10
5
0
−5
−10
−20
−15
−10
−5
0
5
10
15
20
Gene 2 Value

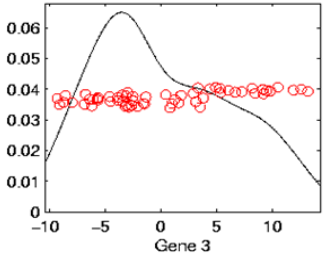
Example



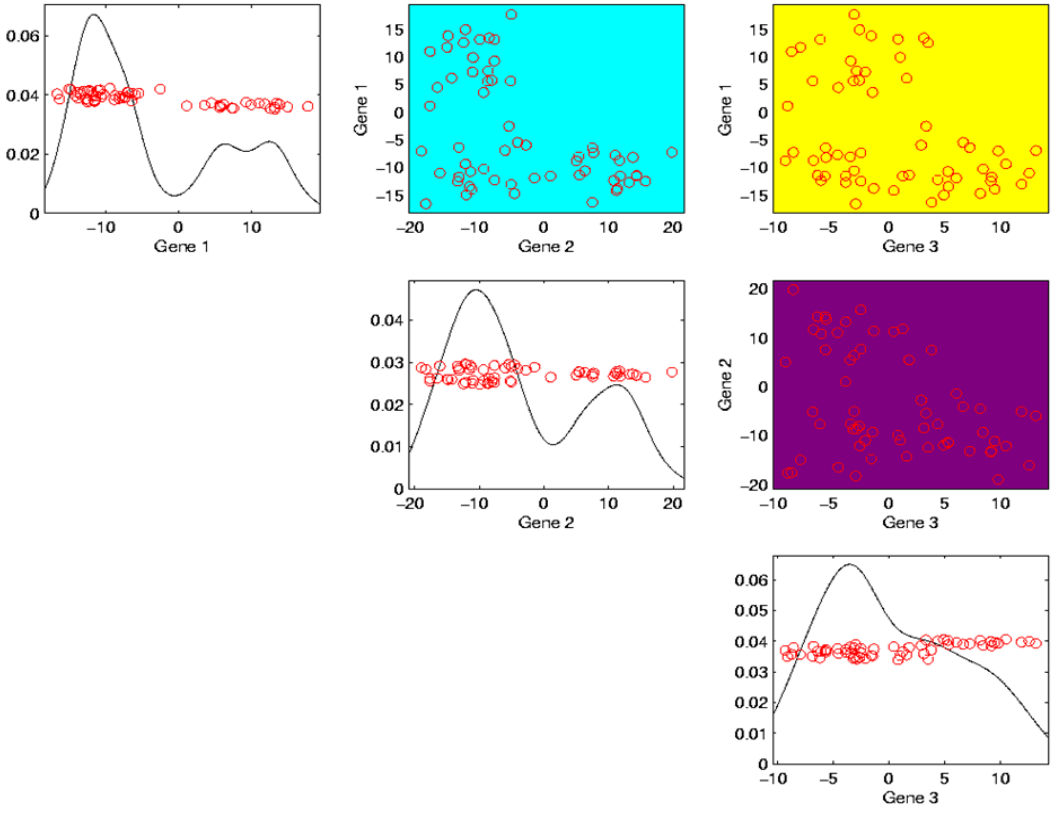
Example


0.06
0.04
0.02
0
−10
0
10
Gene 1

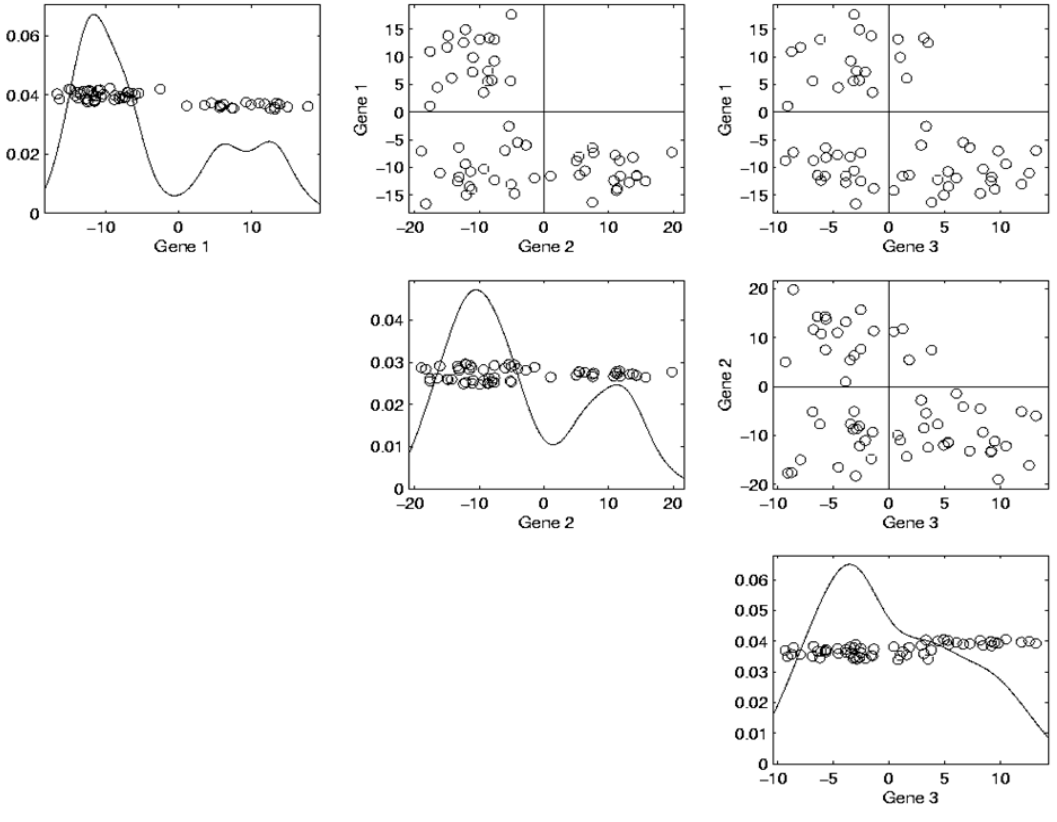

0.04
0.03
0.02
0.01
0
−20
−10
0
10
20
Gene 2


0.06
0.05
0.04
0.03
0.02
0.01
0
−10
−5
0
5
10
Gene 3

Example


0.06
0.04
0.02
0
−10
0
10
Gene 1
15
10
5
0
−5
−10
−15
Gene 1
−20
−10
0
10
20
Gene 2
15
10
5
0
−5
−10
−15
Gene 1
−10
−5
0
5
10
Gene 3
0.04
0.03
0.02
0.01
0
−20
−10
0
10
20
Gene 2
20
10
0
−10
−20
Gene 2
−10
−5
0
5
10
Gene 3
0.06
0.05
0.04
0.03
0.02
0.01
0
−10
−5
0
5
10
Gene 3

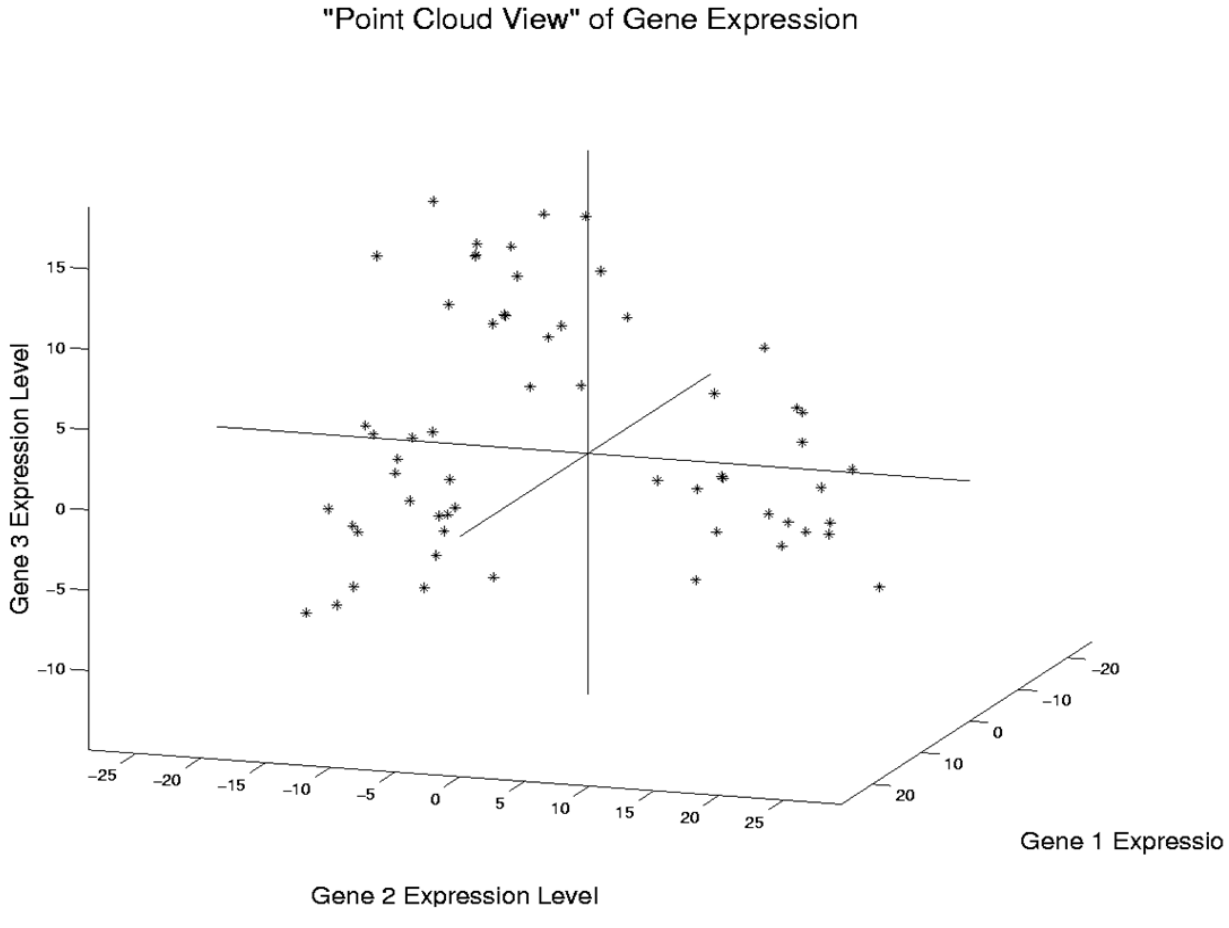
Example: Scatter plot matrix



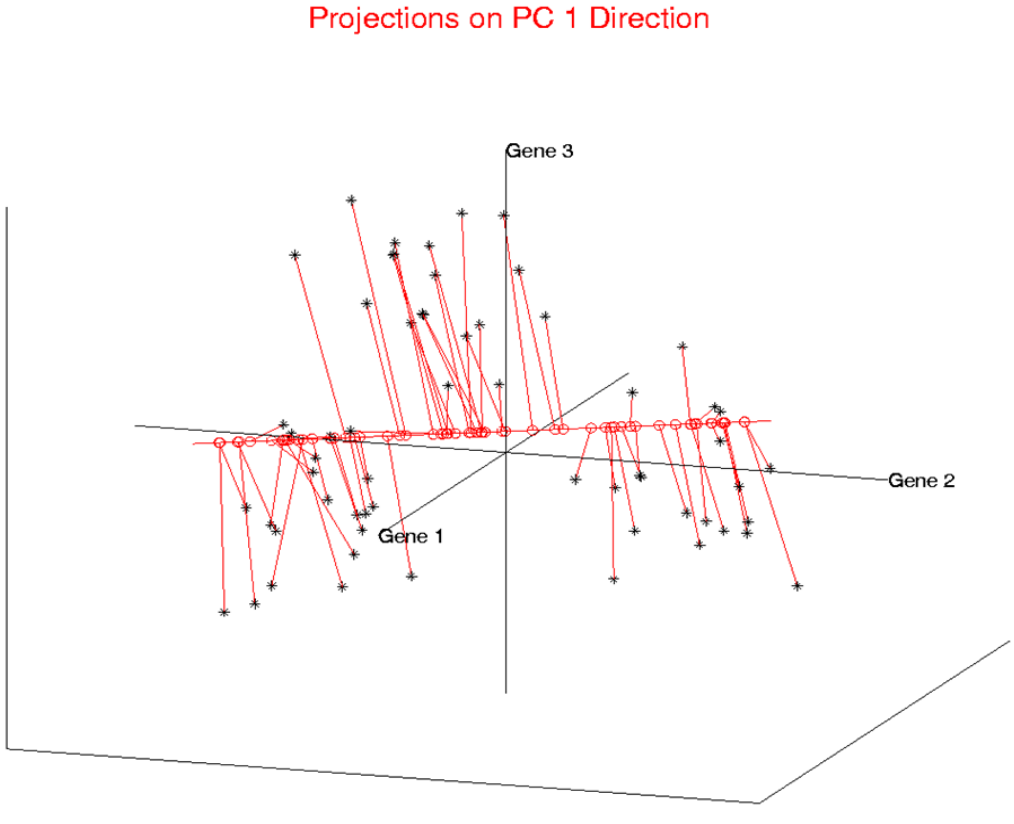
Principal Component Analysis

- Important set of directions: Principal Components
- Find directions of maximal (projected) variation
- Compute sequentially
- On orthogonal subspaces

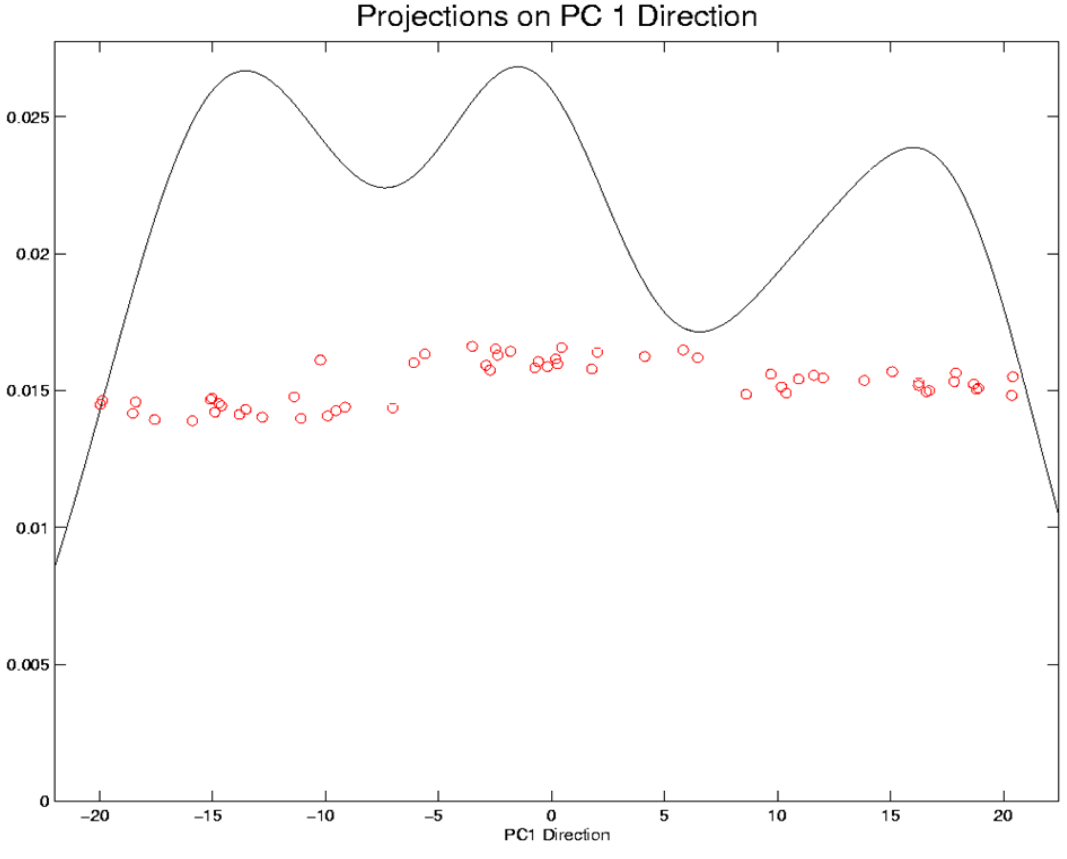
Example


"Point Cloud View" of Gene Expression
Gene 3 Expression Level
15
10
5
0
−5
−10
−25
−20
−15
−10
−5
0
5
10
15
20
25
Gene 2 Expression Level
−20
−10
0
10
20
Gene 1 Expressio

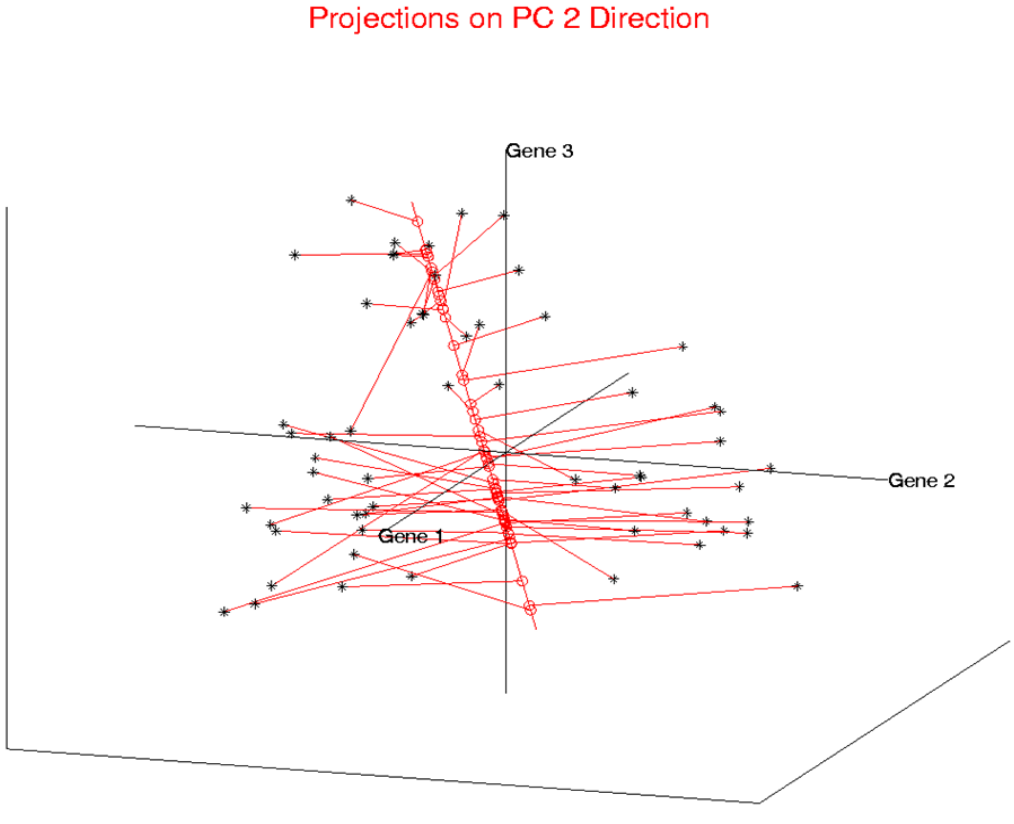
Example



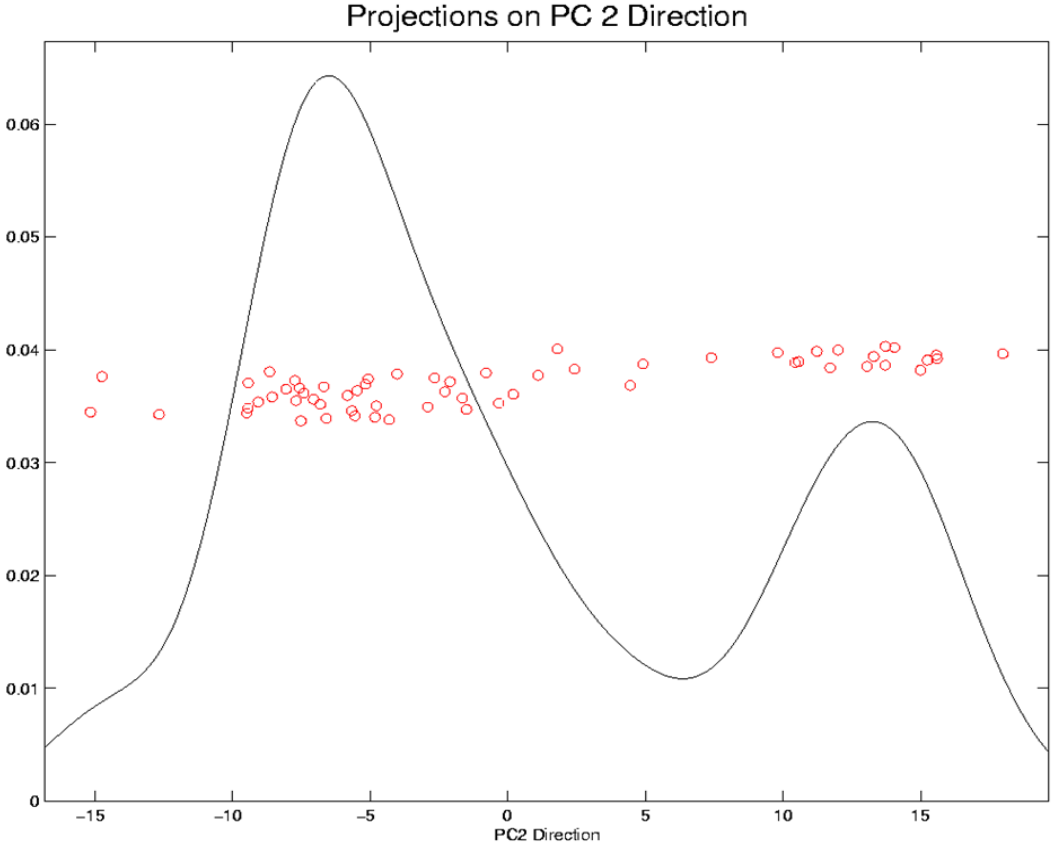
Example


Projections on PC 1 Direction
0.025
0.02
0.015
0.01
0.005
0
−20
−15
−10
−5
0
5
10
15
20
PC1 Direction

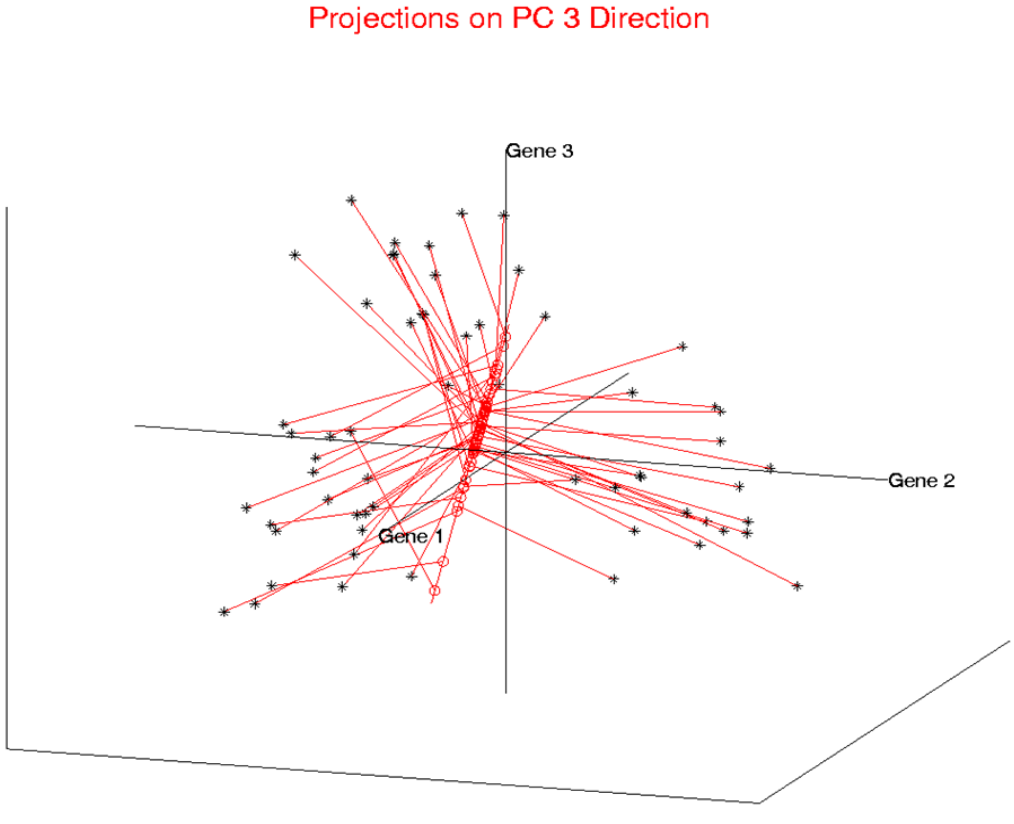
Example



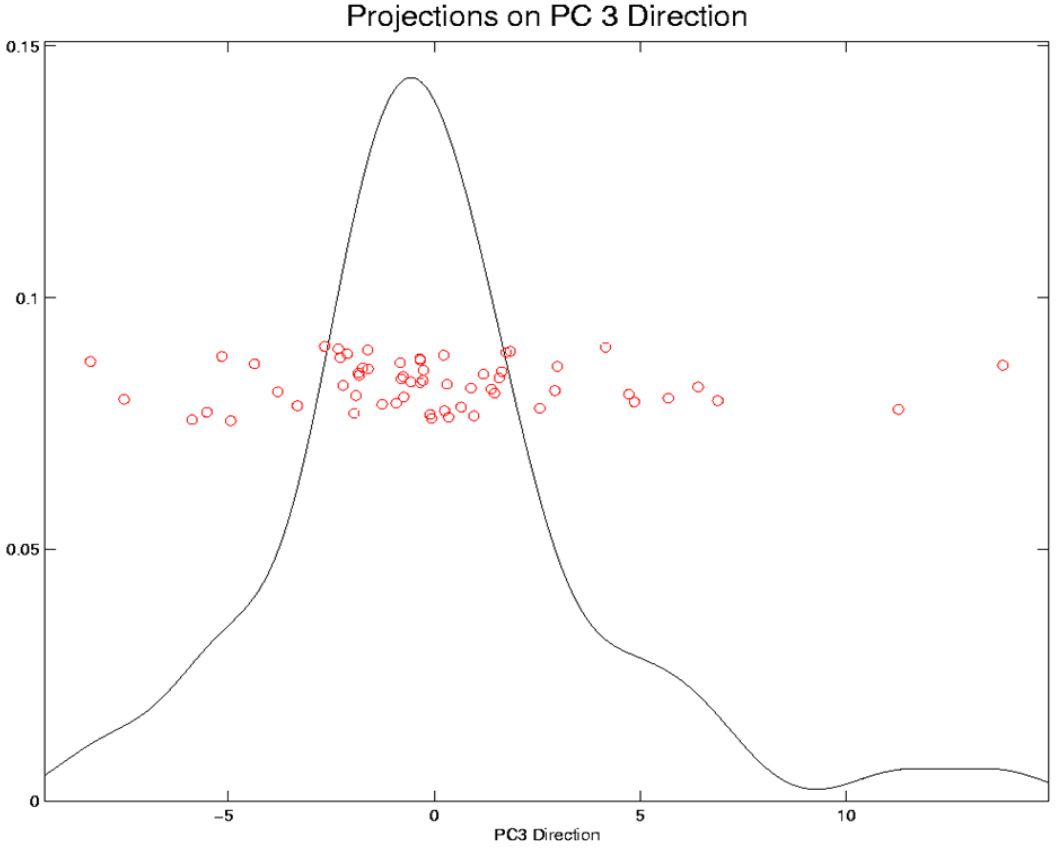
Example


Projections on PC 2 Direction
0.06
0.05
0.04
0.03
0.02
0.01
0
−15
−10
−5
0
5
10
15
PC2 Direction

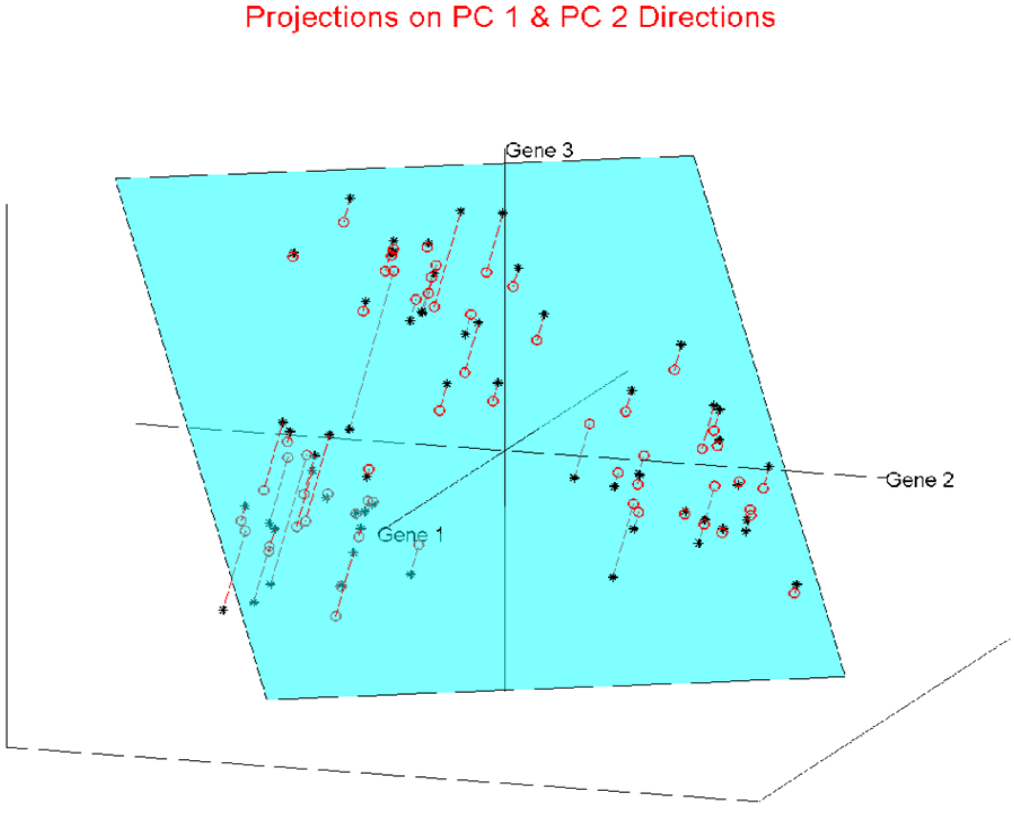
Example



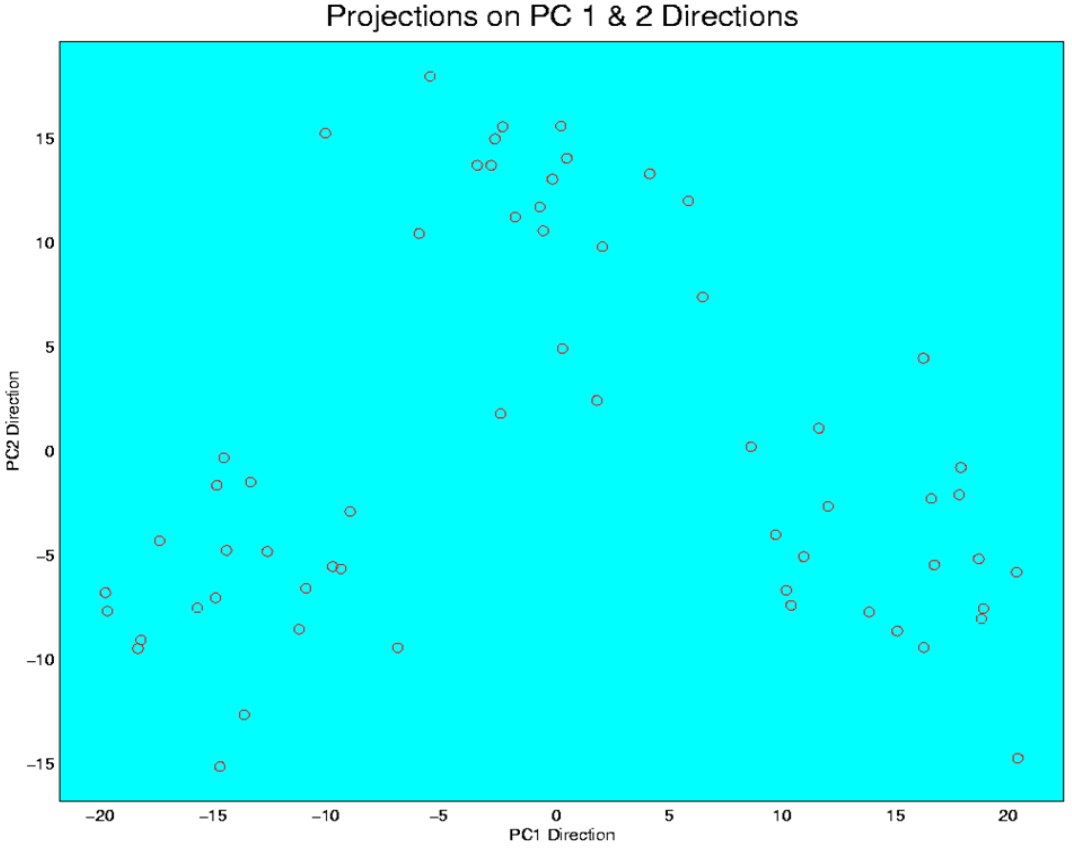
Example


Projections on PC 3 Direction
0.15
0.1
0.05
0
−5
0
5
10
PC3 Direction

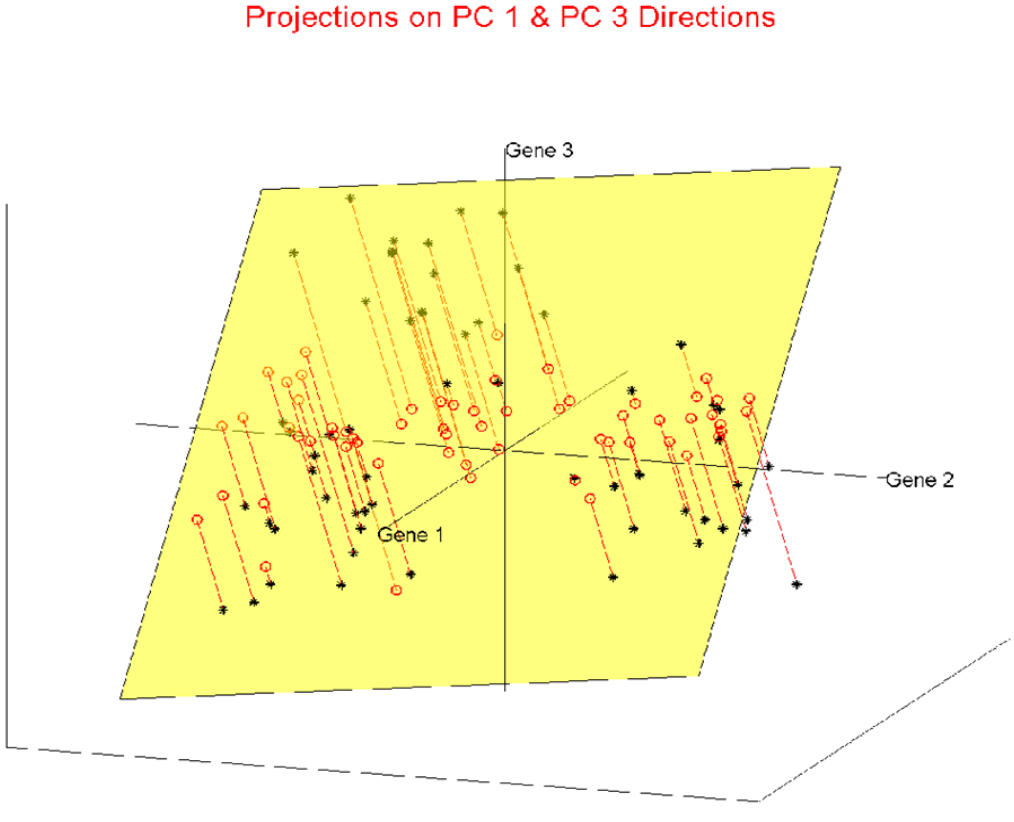
Example



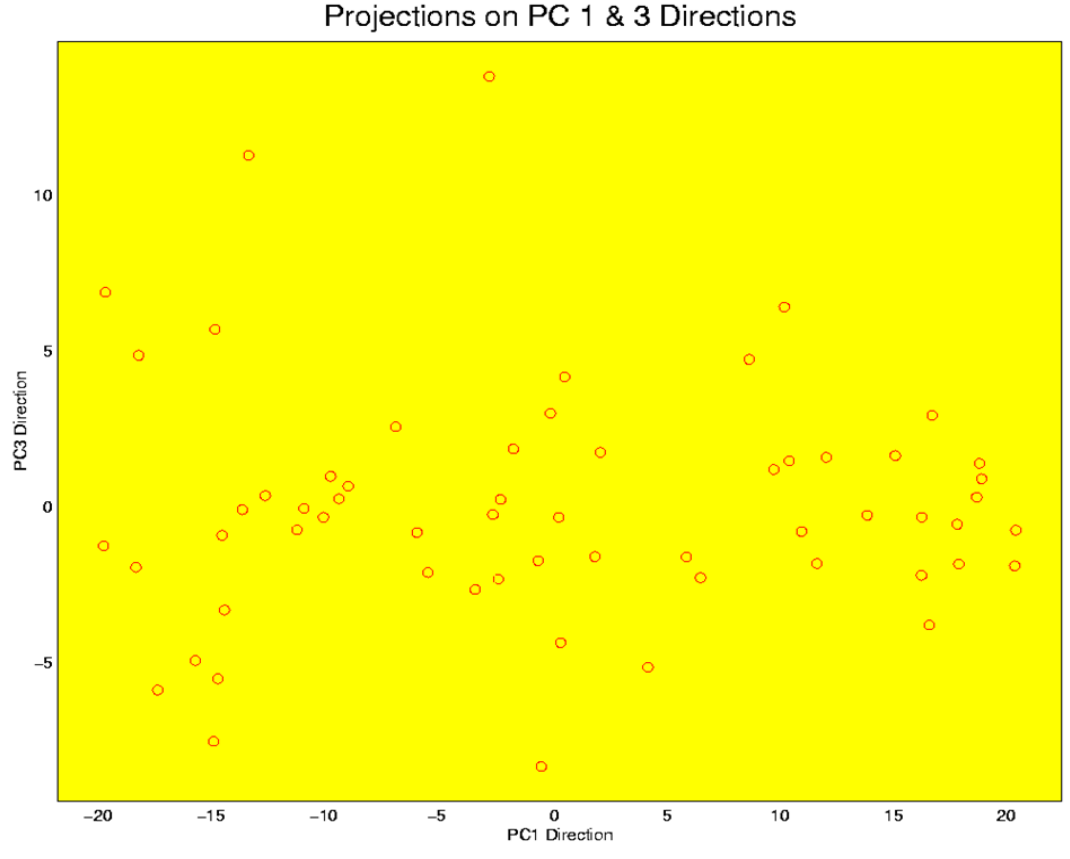
Example


Projections on PC 1 & 2 Directions
PC2 Direction
PC1 Direction
15
10
5
0
−5
−10
−15
−20
−15
−10
−5
0
5
10
15
20

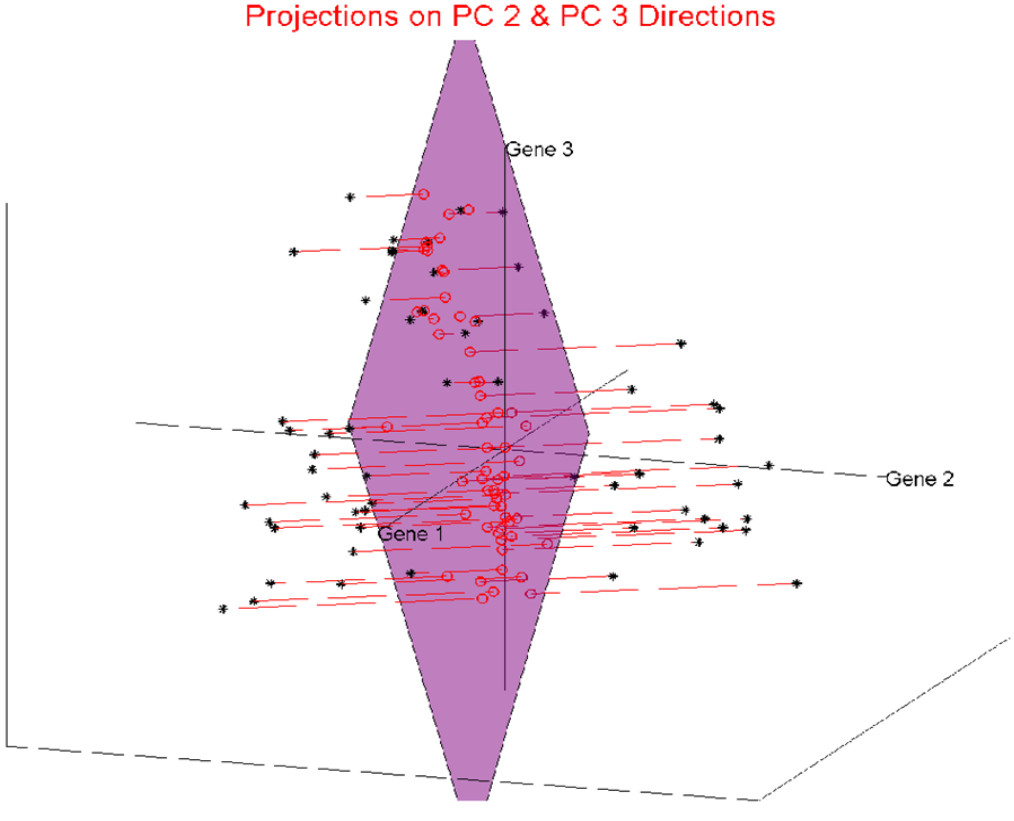
Example



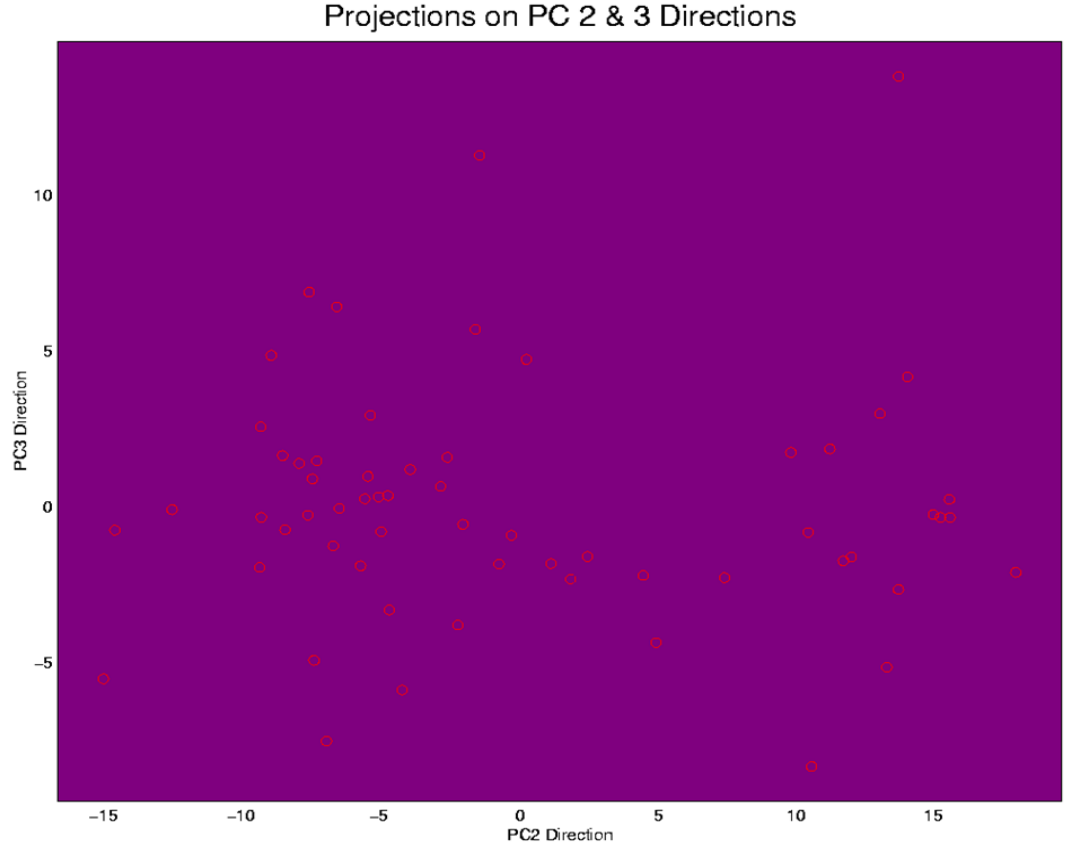
Example


Projections on PC 1 & 3 Directions
PC3 Direction
PC1 Direction
10
5
0
−5
−20
−15
−10
−5
0
5
10
15
20

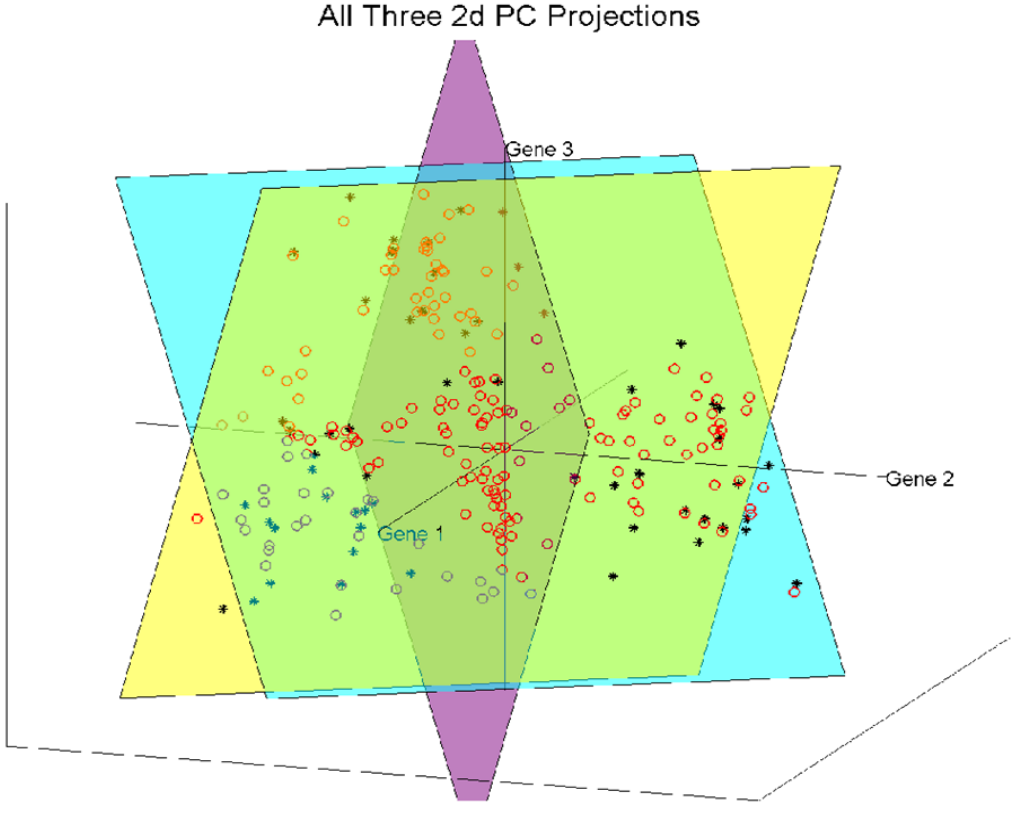
Example



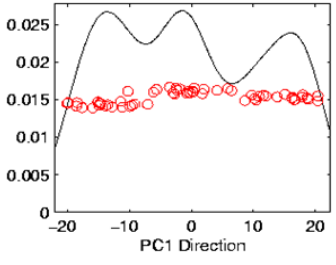
Example

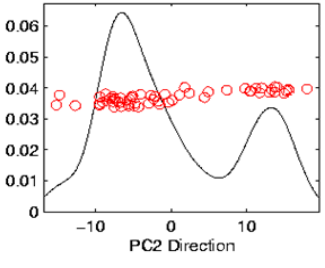

Projections on PC 2 & 3 Directions
PC3 Direction
10
5
0
−5
−15
−10
−5
0
5
10
15
PC2 Direction

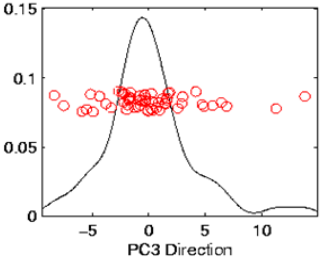
Example



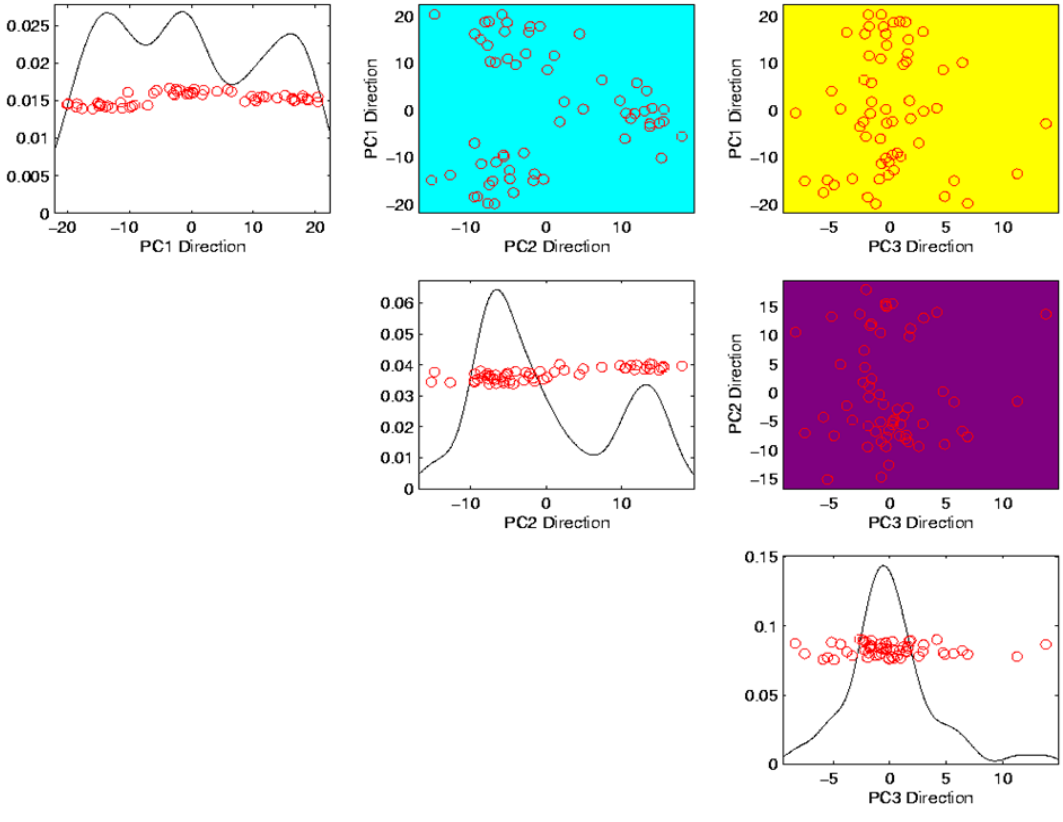
Example


0.025
0.02
0.015
0.01
0.005
0
−20
−10
0
10
20
PC1 Direction

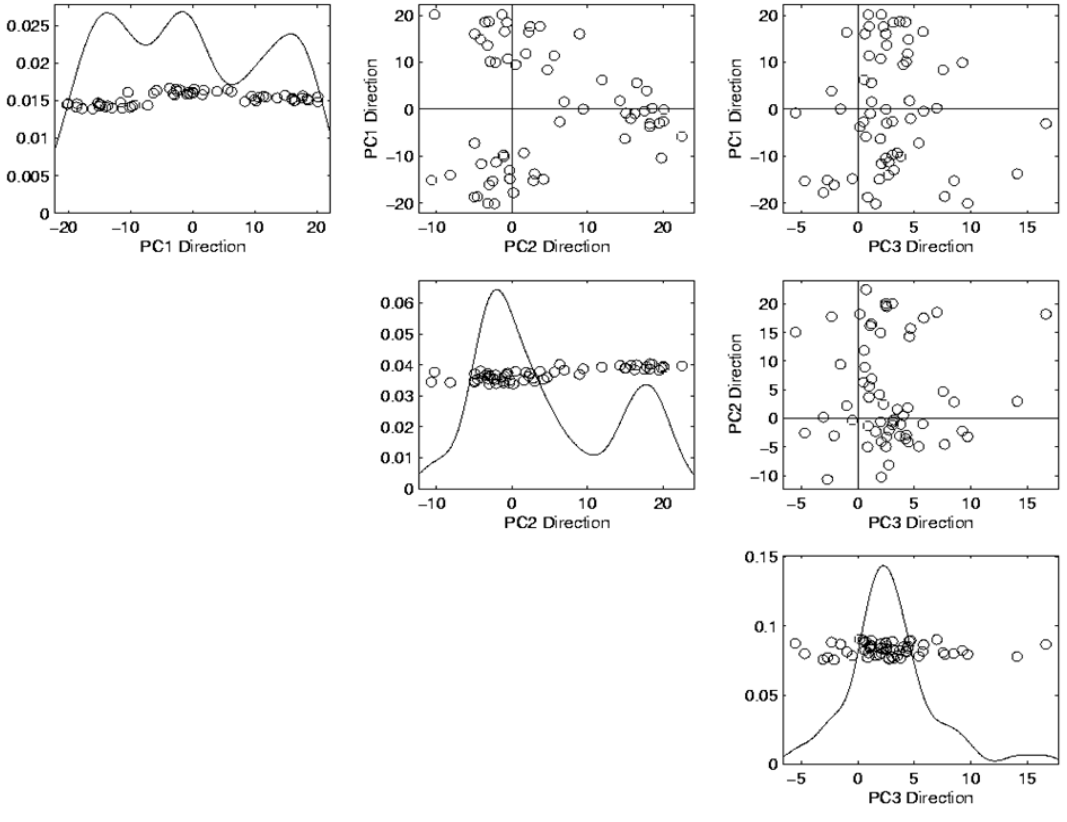

0.06
0.05
0.04
0.03
0.02
0.01
0
−10
0
10
PC2 Direction


0.15
0.1
0.05
0
−5
0
5
10
PC3 Direction

Example

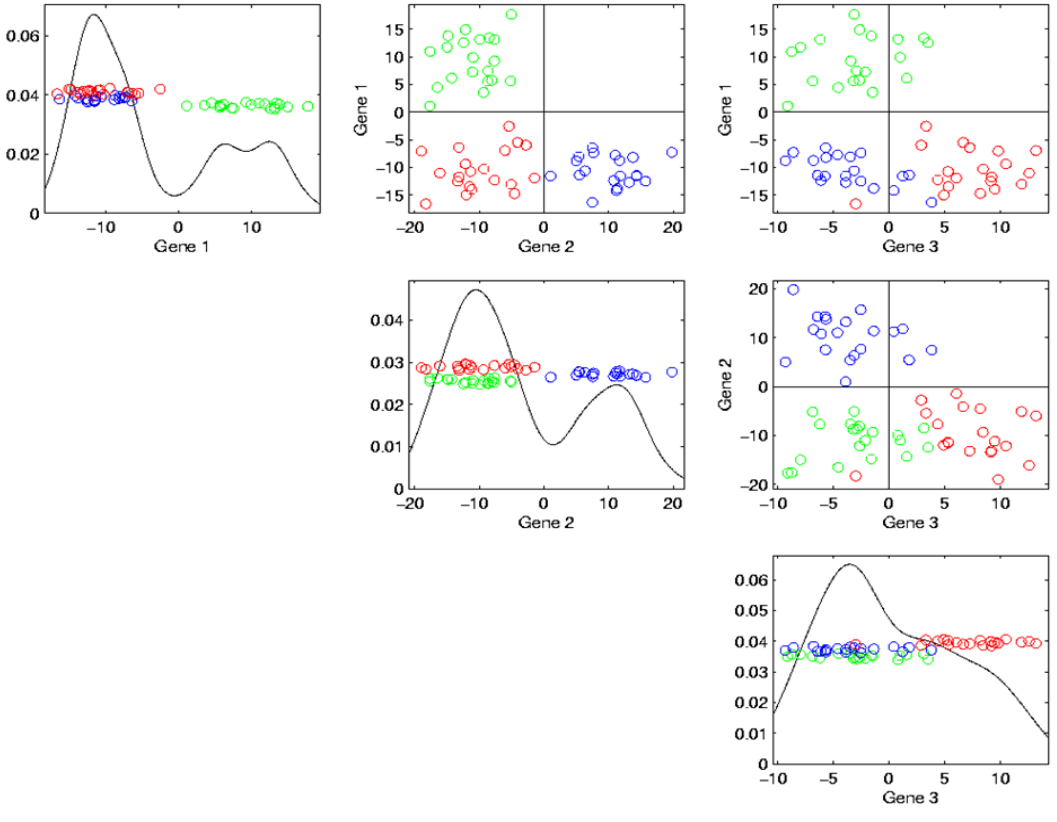

PC1 Direction
PC2 Direction
PC3 Direction
PC1 Direction
PC2 Direction
PC3 Direction
PC2 Direction
PC3 Direction

Example: PCA Scatter plot matrix

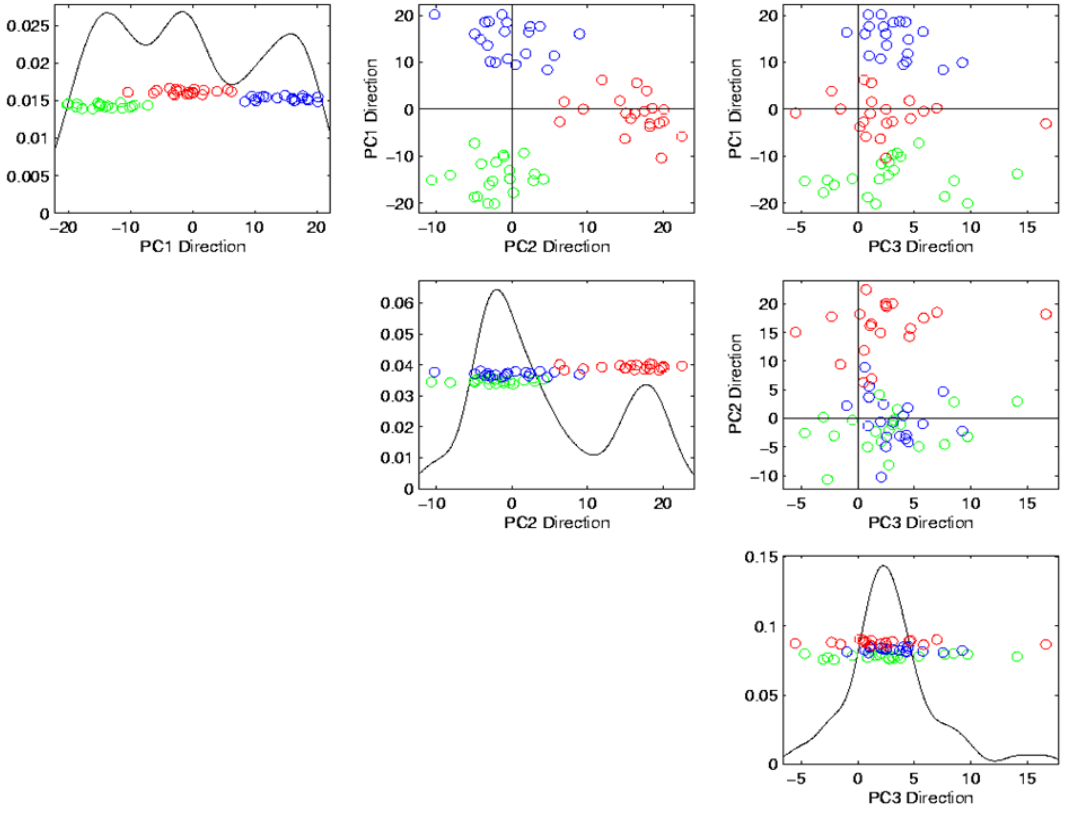


Now highlight 3 clusters

Example: Scatter plot matrix



Example: PCA Scatter plot matrix



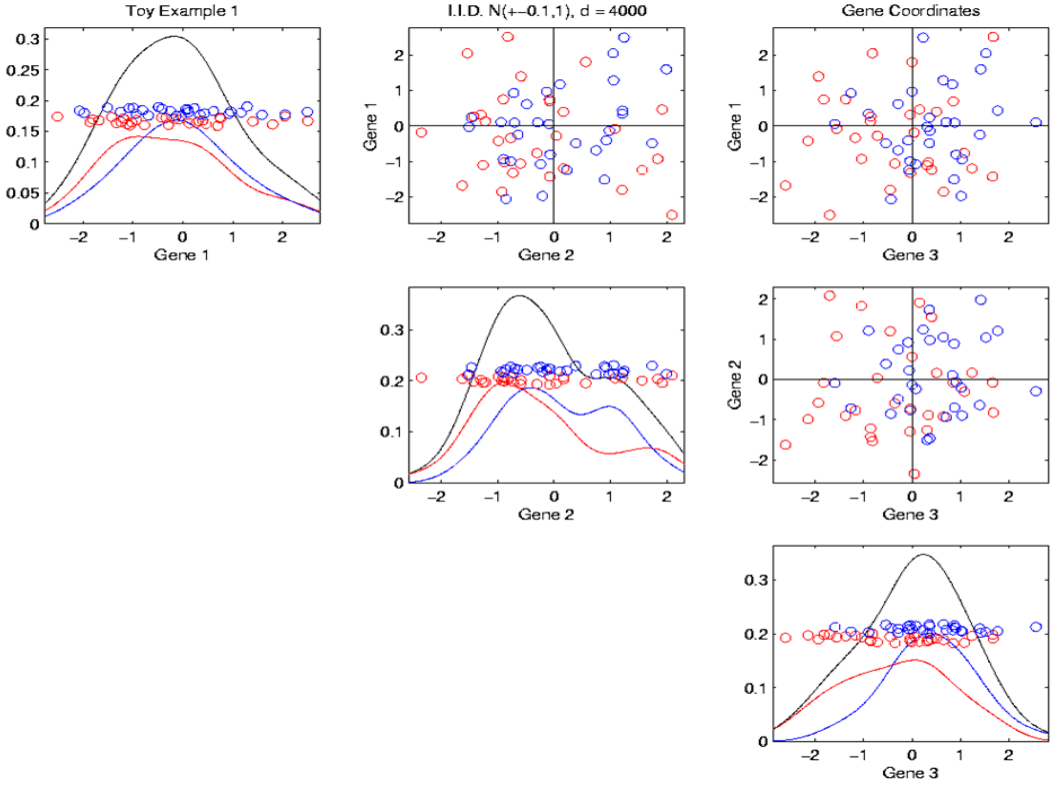
Comparison

- Gene by Gene view: Clusters appear in all 3 scatter plots, but never very separated
- PCA view: 1st PC shows three distinct clusters, better separated than in gene view

Another Example

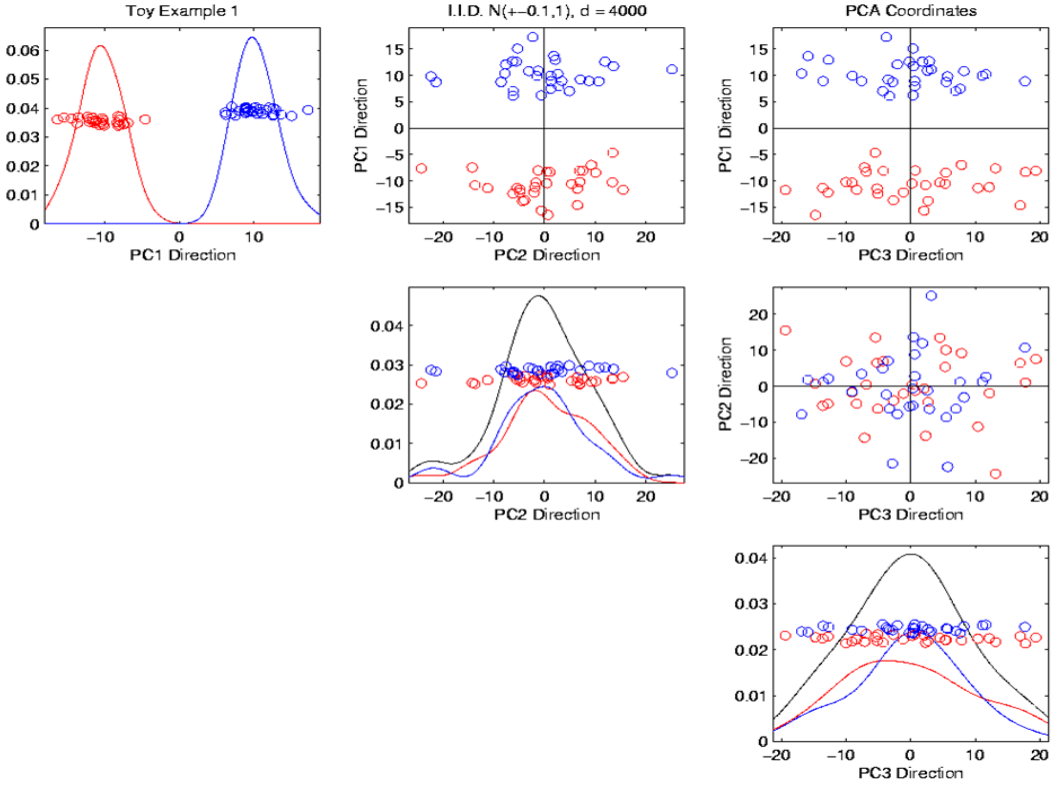
- Now with much higher dimension, # genes = 4000
- 2 clusters with slight mean difference in each variable $N(\pm 0.1, 1)$

Gene by Gene view



2 clusters nearly the same

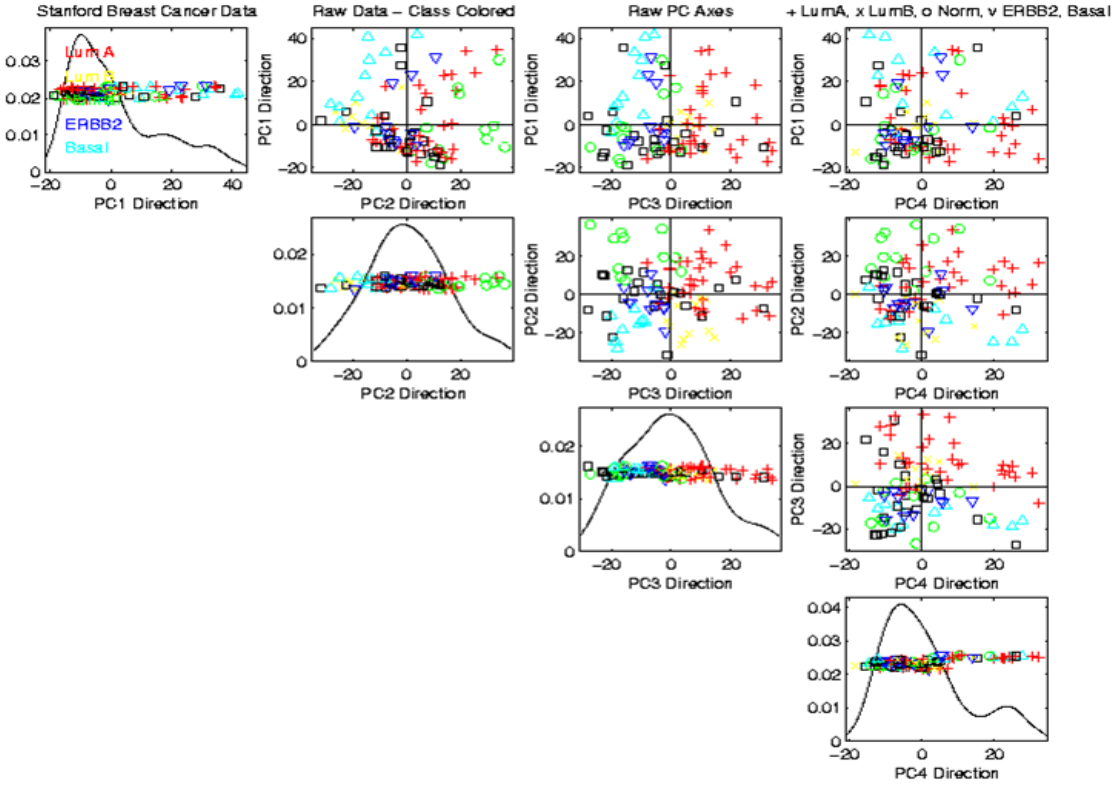
PCA view



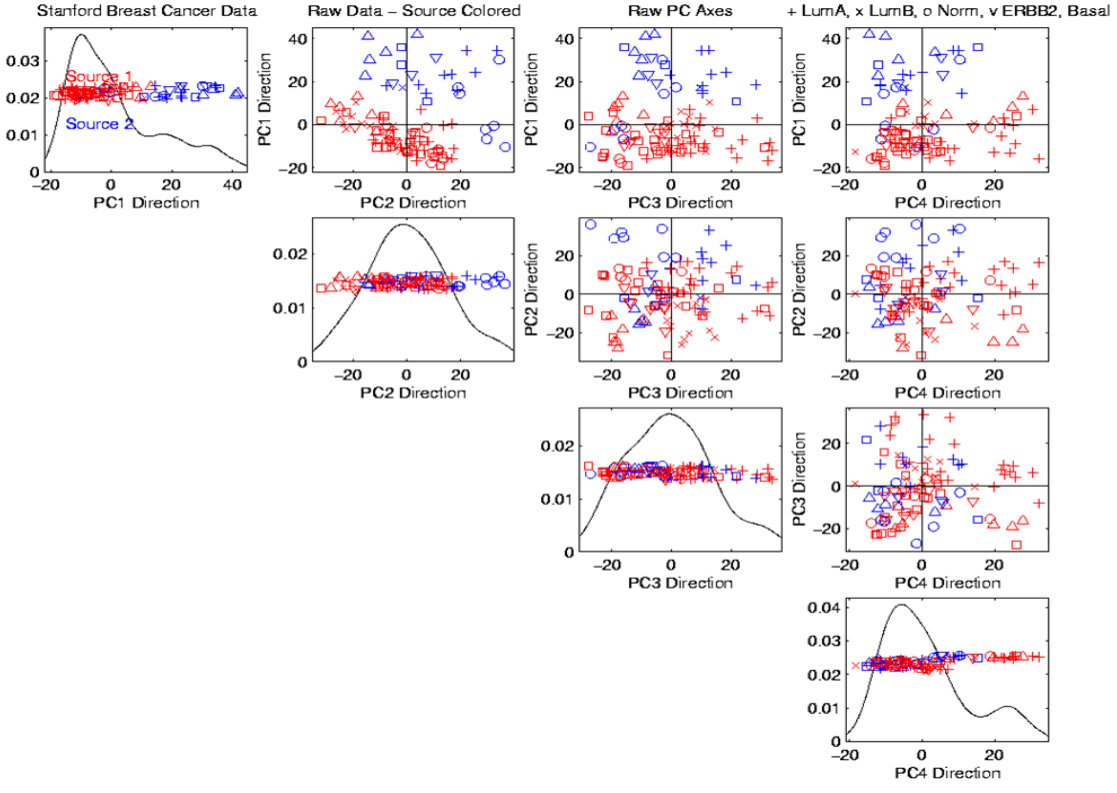
Huge difference in the 1st PC direction

Alternative views can show more

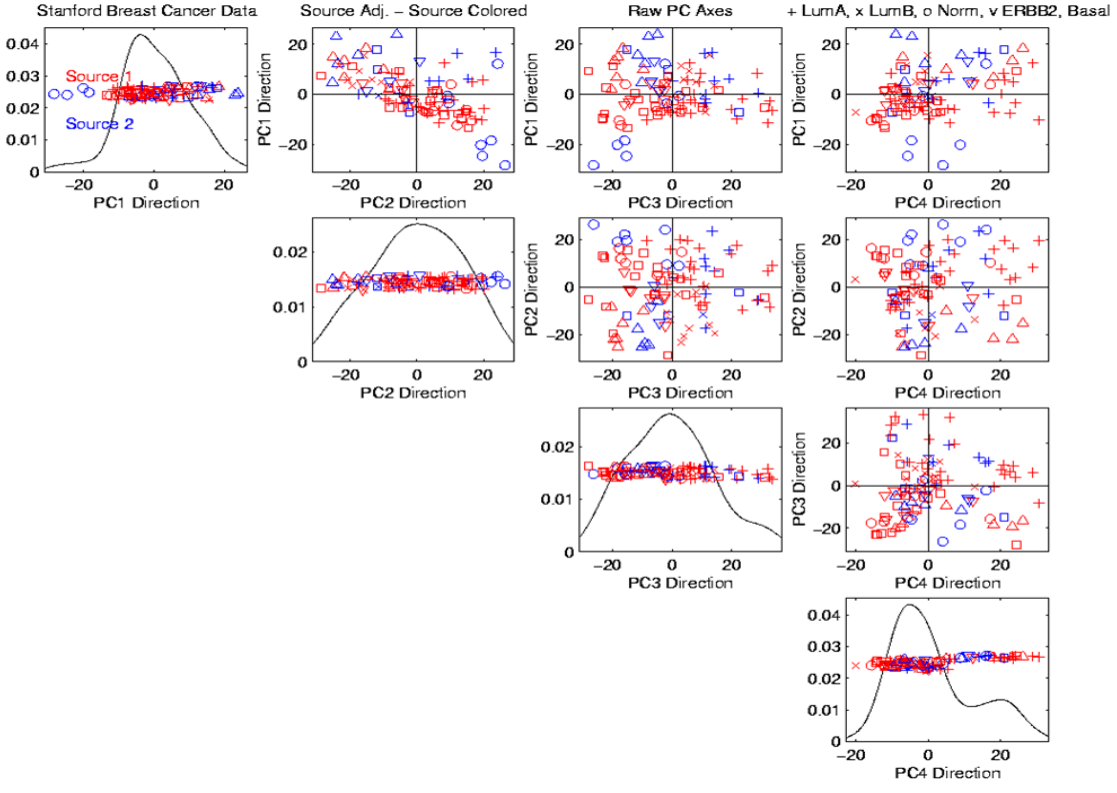
Breast Cancer Data



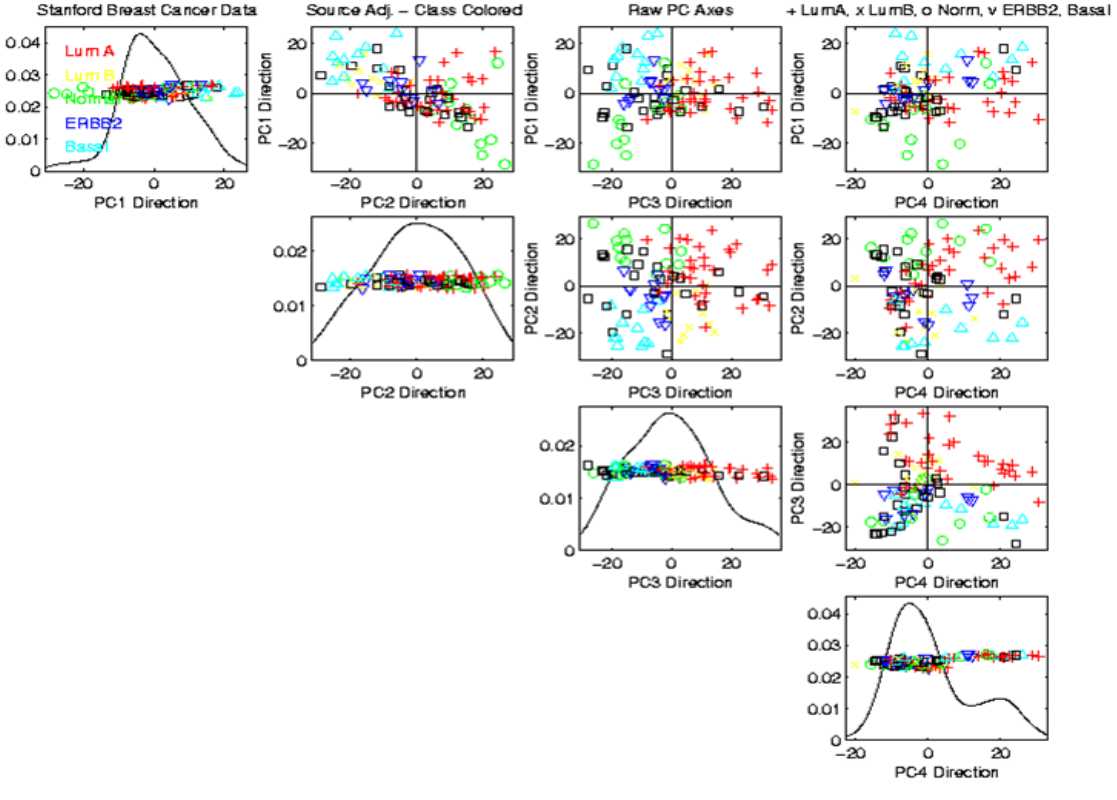
Breast Cancer Data



Breast Cancer Data



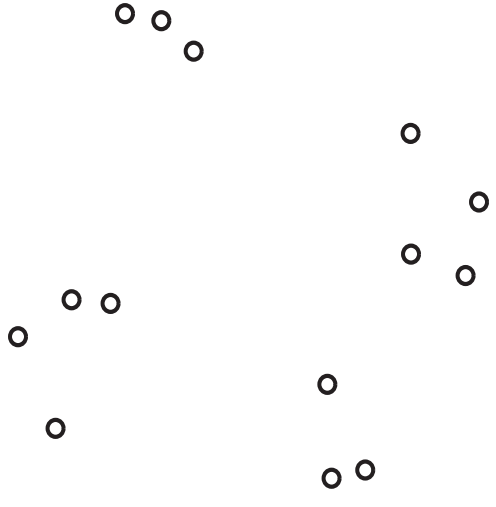
Breast Cancer Data



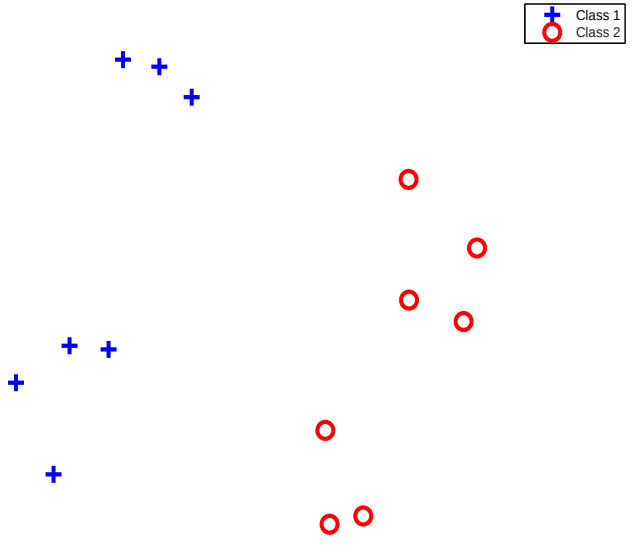
Classification

- Class 1: $\mathbf{X}^1_{d\times m}$, Class 2: $\mathbf{X}^2_{d\times n}$
- Example: microarray from cancer / non-cancer patients.
- Use training data from Class 1 & Class 2.
- Develop a rule for assigning new data to a class label.
- Based on measurements, want to predict if new patients are Healthy or Ill

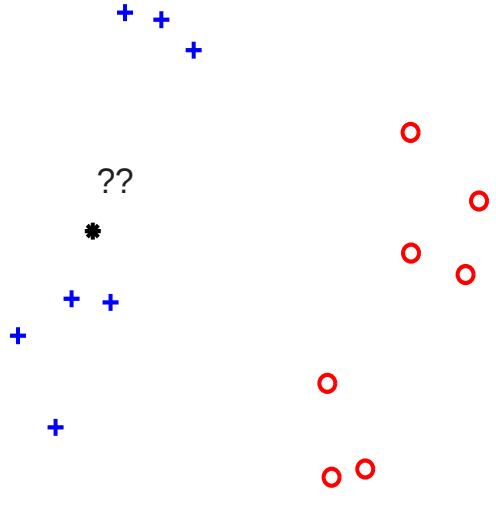
Example



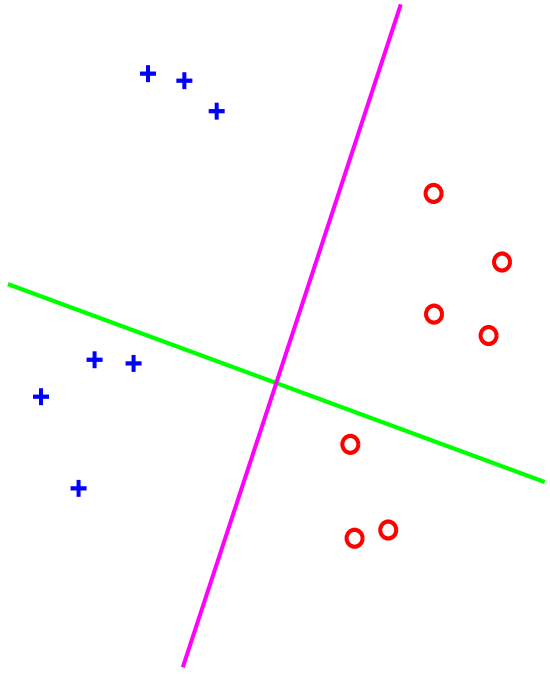
Example



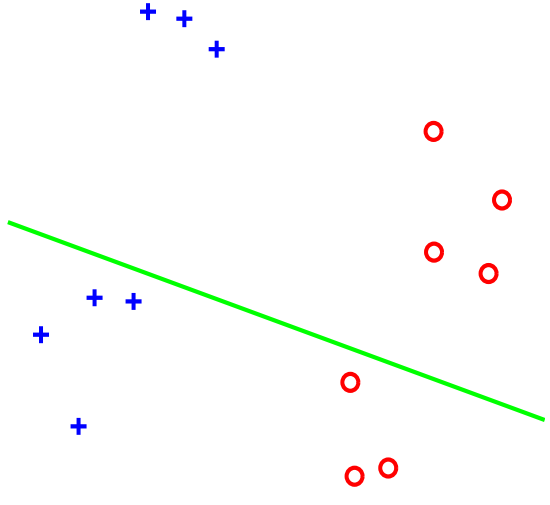
Example



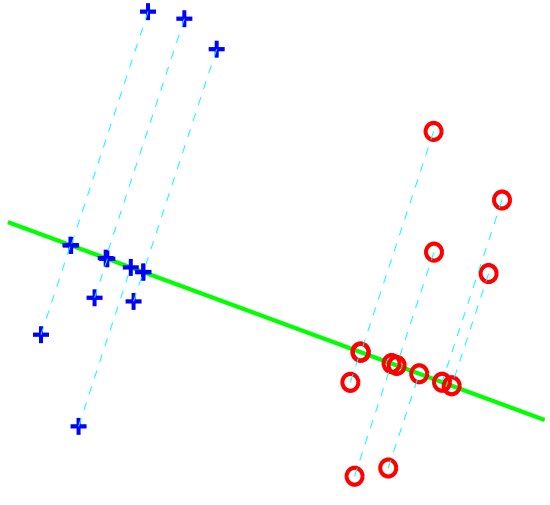
Example



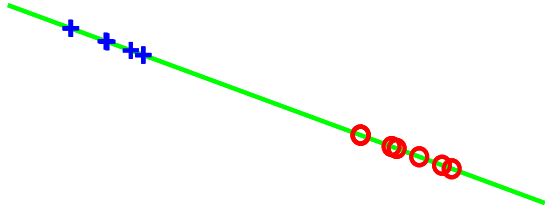
Example



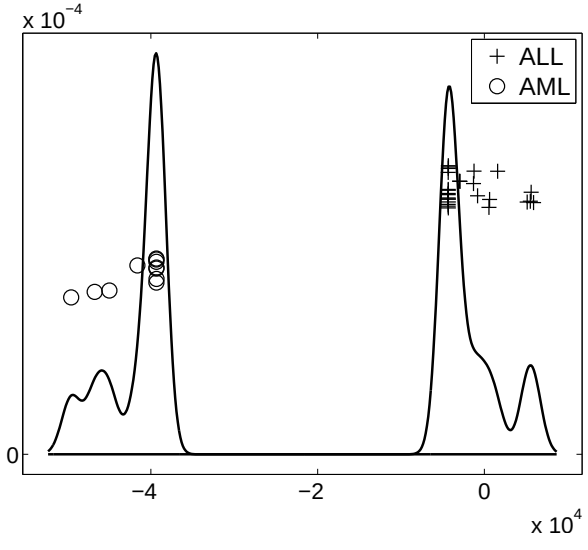
Example



Example



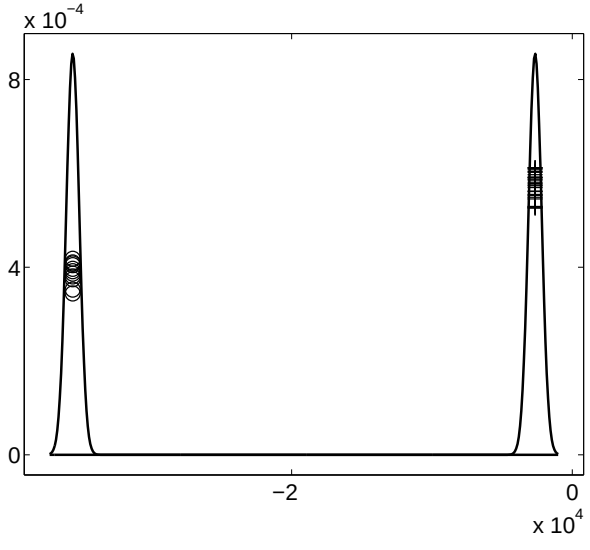
Classification visualization



$n = 38$ and $p = 7129$

Projections of leukemia microarray data onto the SVM direction vector.

Classification visualization



$n = 38$ and $p = 7129$

Projections of leukemia microarray data onto the MDP direction vector (defined only for HDLSS data).

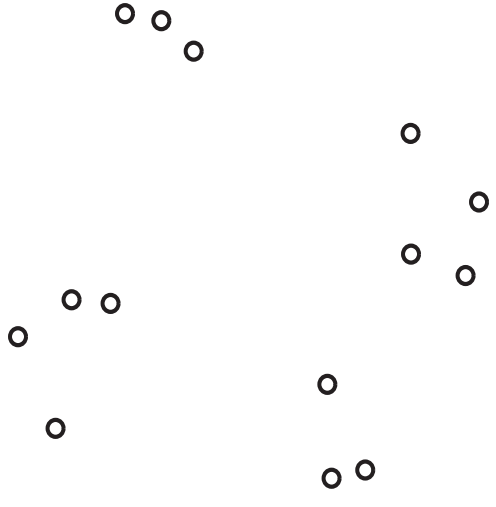
Cluster Analysis

Goal: we want to see if some natural groups (classes, clusters) of individuals exists.

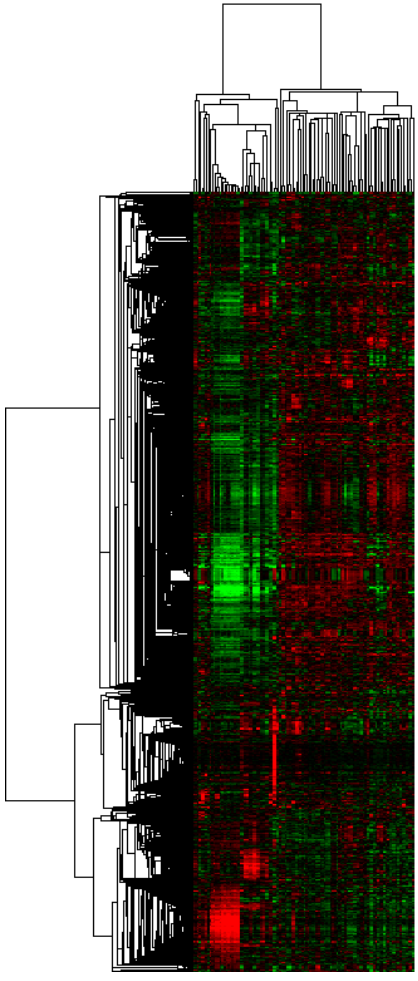
The groups should be as homogeneous as possible and the differences between the groups as large as possible.

- Choice of proximity measure
- Choice of grouping algorithm

Cluster Analysis Example



Example: Ovarian Cancer Data



$n = 108$ and $p = 28,475$

Goal

- Understand why and how multivariate methods are used
- Understand when a multivariate method should be used
- Understand some of the theory behind the methods
- Gain experience applying multivariate methods and interpreting results